Computer-aided diagnosis of prostate cancer using multi-parametric MRI - Evaluation of feature extraction and classification

By

Sean R. Babcock

A Thesis/Dissertation

Presented to the Department of Medical Informatics and Clinical Epidemiology and the Oregon Health & Science University School of Medicine in partial fulfillment of the requirements for the degree of

Master of Science

March 2019

TABLE OF CONTENTS

List of Figures and Tables

List of abbreviations

PCa: prostate cancer
MRI: magnetic resonance imaging
mp-MRI: multiparametric magnetic resonance imaging
PSA: prostate-specific antigen
TRUS: transrectal ultrasound-guided
DRE: digital rectal exam
PZ: Peripheral zone
TZ: Transition zone
CZ: Central zone
AFS: Anterior fibromuscular stroma
T2W: T-2 weighted
T2WI: T-2 weighted image
DWI: Diffusion-weighted image
ADC: apparent diffusion coefficient
ADCmap: apparent diffusion coefficient map
NN: neural network
CNN: convolutional neural network
GLCM: grey-level co-occurrence matrix
GLSZM: grey-level size zone matrix
GLRLM: grey-level run length matrix
LBP: local binary pattern
FD: fractal dimension
MRFD: multi-resolution fractal dimension
ROI: region of interest
2D: two-dimensional
3D: three-dimensional
ROC: receiver operating characteristic
AUC: area under the curve
DAE: de-noising autoencoder
VAE: variational autoencoder
CAE: convolutional autoencoder
ReLU: rectified linear unit
FC: fully connected
SRE: Short Run Emphasis
LRE: Long Run Emphasis
GLN: Grey Level Non-uniformity
RLN: Run Length Non-uniformity
RP: Run percentage
LGRE: Low Grey Level Run Emphasis
HGRE: High Grey Level Run Emphasis
SRLGE: Short Run Low Grey Level Emphasis
SRHGE: Short Run High Grey Level Emphasis
LRLGE: Long Run Low Grey Level Emphasis

LRHGE: Long Run High Grey Level Emphasis
DBC: Differential Box Counting
mRMR: minimum-redundancy–maximum-relevancy
SVM: support vector machine
RF: random forest
MSE: mean squared error
SD: standard deviation
PPV: positive prediction value
NPV: negative prediction value

Acknowledgements

I would like to first thank my thesis advisor Xubo Song PhD. Not only was she my thesis advisor, she was also the instructor in both my machine learning and image analysis classes, which played a key role in my understanding of the concepts needed to complete this thesis. When deciding a thesis topic, she helped drive my interest in computer image analysis, and specifically, computer-aided analysis of cancer images to aid radiologists in early detection and diagnosis. Dr. Song also was key in identifying a dataset and helped me to build a useful, and masters thesis level project around this dataset.

I would like to thank the other members of my thesis advisory committee, Ted Laderas PhD, and Guillaume Thibault PhD, for their expertise, guidance, and insight into the technical aspects of this project and the thesis dissertation process as a whole. I had the pleasure of knowing Ted from many of the classes I took during my time in the DMICE BCB program. Ted offered a unique insight into the bioinformatics and clinical aspects of this project, which were extremely helpful during my work. Guillaume brought a wealth of technical expertise in computer image analysis. In addition to currently working in the field of computer image analysis pertaining to cancer research, he has also published several papers on the subject.

Without several key contributors to this project, I would have never had the resources to succeed in my goals. First, Dr. Fergus Coakley M.B. B.Ch who was generous enough to provide the image dataset used in my thesis study. Dr. Coakley also offered radiologist resources to compile the dataset and offered understanding to how the dataset was classified. Secondly, I would like to offer a special thanks to Archana Machireddy, a PhD student in the Center for Spoken Language Understanding who first worked with a radiologist to collate the dataset and put it into a form which I was able to apply to my thesis goals. Without these two individuals, the project would have never gotten off the ground.

My family, Bob, Diane, and Darcie Babcock, played a critical role in not only supporting my decision to enter the bioinformatics and computational biomedicine program enabling me to achieve all I have this far. By also offering their support and understanding during the difficult times, as well as sharing their joy of my accomplishments, they gave me the will to move forward. I would specially like to thank my sister, Darcie Babcock, who has worked at OHSU for many years and was responsible for getting me interested in the bioinformatics field by requesting my help using sequencing software in a genetics project. From this first kernel of understanding I decided to find out more about bioinformatics and enroll in the DMICE BCB program.

Lastly with all my heart I wish to thank my late wife Jacqueline Babcock. Although she is no longer with me, she has been a constant source of strength and guidance in all my endeavors, including the desire to study and treat cancer and hopefully one day be involved in finding promising cures and treatments.

Abstract

According to the NIH[1], prostate cancer is the third leading cancer type in the United States. It is projected in 2018 that prostate cancer (PCa) will account for the most newly diagnosed cancer cases (164,690), which accounts for 9.5% of all new cancer cases, and be the second leading cause of death (29,430) among males in the United States[1, 2]. Two factors determine the 5-year survival rate of prostate cancer: early diagnosis and tumor localization in the prostate gland[1]. To underscore this point, the 5-year survival rate for localized prostate cancer diagnosed early is 100%[1]. Current methods of prostate cancer detection can result in overdiagnosis and overtreatment of non-aggressive cancer by failing to distinguish between non-aggressive cancer and more aggressive cancer (Gleason score $\geq$ 7)[3]. In addition, variability between radiologists when reading and grading MR images is also a cause of misdiagnosis and overtreatment. Overdiagnosis and overtreatment can be costly, and put unnecessary burden on patients, insurance, and medical facilities. Imaging has increasingly shown promise in aiding in the detection and diagnosis of prostate cancer. Specifically, multiparametric MRI (mp-MRI) has shown great promise in developing computer-aided techniques to aid radiologists in detecting and accurately assessing prostate cancer[4]. In fact, computer imaging studies from Stanford University in 2017 using deep learning methods, demonstrated that for skin cancer diagnosis the deep learning classification method matched dermatologist accuracy[5] and for pneumonia diagnosis, the deep learning image analysis method exceeded radiologist diagnosis accuracy[6].

The goal of this project was to evaluate the feasibility of machine learning methods based on feature extraction and classification for differentiating benign and malignant prostate cancer tumors in mp-MRI with the intent of building an algorithm that best reduces over-biopsy and enables in-silico biopsy. The model used regions of interest (ROI) in 2D multiparametric magnetic resonance images (mp-MRI) selected by the radiologist as possible prostate cancer tumors. The specific aim of this project was to evaluate various extracted texture features, dimensionality reduction, and machine learning classification methods to determine the computer-aided analysis model that provides the greatest classification accuracy for malignant and benign prostate cancer images. From these image level models, a patient level diagnosis decision support model was determined that best diagnoses malignant cancer patients.

This project has demonstrated several key results. First, I have shown the viability of my algorithm using image feature extraction and classification to develop a computer-aided clinical diagnosis decision support model, which has shown perfect malignant patient diagnosis using the dataset available to me for this study. Secondly, I have shown that mp-MRI is the preferred imaging method than that of using single MRI image types such as T-2 weighted (T2W) or apparent diffusion coefficient maps (ADCmaps) for this type of model.

Chapter 1: Introduction

Background

As of 2015, the current model for prostate cancer screening and detection is the analysis of prostate-specific antigen (PSA) levels and/or a digital rectal exam (DRE)[7, 8]. If indicated by elevated PSA levels or DRE screening, a transrectal ultrasound-guided (TRUS) biopsy or MRI guided biopsy are used as a follow up diagnosis step[7, 8]. Biopsy has been shown to be the most accurate method of confirming the presence of prostate cancer, however PSA testing, which prompts a biopsy, has shown to be unreliable in finding a suitable threshold to indicate if a biopsy is required[9]. While the PSA screen and TRUS biopsy diagnostic methods are reported to have increased the detection of prostate cancer, they suffer from overdiagnosis and overtreatment due to many of the identified cases being low risk and clinically insignificant, and also may not detect anterior tumors that are of a small size[7]. A conventional biopsy taken without imaging guidance, known as a "blind" biopsy, have been reported to find more non-aggressive cancer than aggressive cancer[3]. Another prostate cancer detection method, the DRE, is also shown to be effective in detecting prostate cancer in the posterior peripheral zone (PZ), but is ineffective in detecting cancer in all others zones of prostate gland[10]. Figure 1[9] shows the zones of the prostate gland with the PZ highlighted in pink.

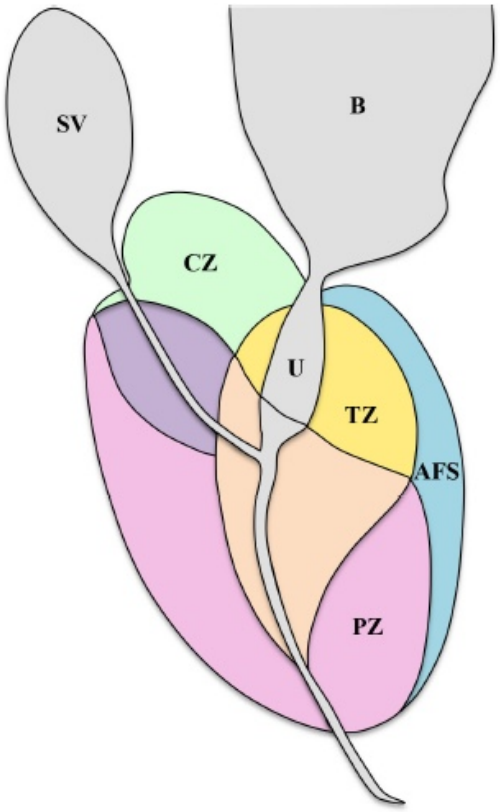

Figure 1. McNeal zonal view of the prostate gland[9]. Central zone: CZ (green), Transition zone: TZ (yellow), Anterior fibromuscular stroma: AFS (blue), Peripheral zone: PZ (pink), Seminal vesicle: SV (grey), Bladder: B (grey), Urethra: U (grey).

Regardless of the biopsy method, the biopsy result is reported with a Gleason score,[11,12] used by pathologists and clinicians to evaluate the current cancer state and aggressiveness, and also to guide clinical treatment options. The Gleason scoring system grades prostate tissue and cells on a scale from 1-5 and represents the biological behavior of the cancer. A grade of 1 is assigned to cancerous tissue that highly resembles healthy prostate tissue; a grade of 5 assigned to cancer cells that look highly abnormal and show cancer-like growth. The final Gleason score report (Gleason sum), is a combination of two scores having a potential maximum score of 10, for example, 7=4+3. The two scores in the Gleason sum represent two areas of the prostate tumor, where the first score represents the majority (largest area) cancer grade, and the second score represents the minority (smaller area) cancer grade.

Recently, the use of imaging is an emerging method for prostate cancer detection and evaluation, where the use of Magnetic Resonance Imaging (MRI), specifically multiparametric MRI (mp-MRI), is gaining popularity[4,7,9,13]. Computer image analysis, using machine learning techniques, offers a new path in prostate cancer detection and diagnosis by overcoming deficiencies that exist in current prostate cancer detection and diagnosis methodologies. Currently, humans are involved in the analysis and interpretation of images in prostate cancer detection and diagnosis. Limitations and issues facing a radiologist can be divided into three areas: human visual perception, the complexity and overlap of tumor features in the MR images being analyzed, and inter/intra-human variability[3,4]. Image complexities further hinder the limitations on human visual perceptions due to the complex nature of patterns, intensity variations, and structures seen in many cancer tumors[4]. In addition, reading and interpreting MR images is a subjective process and is highly influenced by a radiologist's experience level, training, and a lack of standardization in image interpretation[14]. The subjective nature of image interpretation can lead to errors and low repeatability in the diagnosis of prostate cancer, causing costly and invasive over biopsy with possible incorrect or overtreatment, and, in the extreme case, undertreatment of a patient with a malignant tumor. To underscore experience level as an issue, several image analysis studies using deep learning methods from Stanford University in 2017, demonstrated that for skin cancer diagnosis, the deep learning classification method matched dermatologist accuracy[5] and for pneumonia diagnosis, the deep learning CheXNet image analysis algorithm exceeded 4 Stanford radiologists diagnosis accuracy[6].

To date, many of the computer image analysis efforts based on previous ground truth biopsy results in prostate cancer detection have been directed at aiding radiologists by removing many of the limitations that currently exist in the interpretation of prostate cancer MR images[4,15]. These computer image analysis models also promise to improve cancer detection and treatment. This statement has been verified in several studies. The first, Chan et al.[16], and verified by Dean et al.[17], show a 4% increase in breast cancer detection over current non-computer image analysis assisted methods. In fact, Chan et al.[16] also proposed that a computer image analysis could be of added benefit to radiologists with less training and experience

than more experienced radiologists. Like authors mentioned in the skin cancer study[5], and the pneumonia study[6], the proposed hypothesis that computer image analysis could be of added benefit to radiologists with less training and experience by Chan et al. [16], was later proven in a study of prostate cancer by Hambrock et al.[18]. Several other studies supporting the hypothesis that computer image analysis will improve cancer detection rates are also cited for lung and colon cancers in Lemaitre et al[15].

Beyond aiding pathologists in image interpretation and cancer diagnosis accuracy, computer image analysis of MRIs also shows promise in the areas of in-silico biopsy and radiogenomics. A very accurate, proven, and trusted image analysis method could one day fulfill the promise of in-silico (computer) biopsy, eliminating the need for expensive, invasive, and time-consuming biopsy methods currently used today. Radiogenomics, also referred to as imaging genomics, has recently emerged as a new method in cancer research[19]. By combining image feature analysis and high-throughput genomics, radiogenomics seeks to further the understanding of the biology of cancer tumors, improve diagnosis and treatment selection, and evaluate clinical outcome and response to cancer treatment.

MRI

Multiparametric MRI (mp-MRI) refers to a combination of anatomic imaging (T2-weighted images: T2WI), functional imaging (diffusion weighted images: DWI and dynamic contrast enhancement images: DCE), and metabolic imaging also known as MR spectroscopic images[20]. For this study, the mp-MRI image set will include T2WI and DWI.

Here some basic MRI physics will be explained, and how T2W and DWI relate to prostate cancer imaging. In an MRI magnetic field, a radio frequency (RF) pulse is applied causing the protons in the atoms in the body to spin transverse to the magnetic plane. T2W images are characterized by transverse relaxation time, which is the time it takes protons that have been excited by the RF pulse to decay to their natural state or lose phase with each other. This relaxation time is dependent on the tissue type being imaged with T2W images highlighting both fatty tissue and tissue water content. For prostate cancer, T2W images have shown to be most useful for showing extracapsular extension of cancer tumors and intrusion into the seminal vesicle[13]. However, in the prostate central and transitional zones, where both healthy and cancerous tissue both have similar signal intensity, T2W suffers in its ability to differentiate these tissue types[21]. The DWI functional MRI relies on the diffusion, or motion, of a water molecule in body tissue. More specifically, the cellular density (cellularity) and the degree to which the cellular membranes are intact, will affect the diffusion of the water molecules measured in DWI. In addition to DWI, the apparent diffusion coefficient (ADC) map is also used in prostate cancer imaging. The ADC is a quantitative measure of the amount of diffusion in a tissue. DWI can overcome some of the failings of T2W to detect prostate cancer tumors by its ability to use ADC values to better differentiate between healthy tissue and malignant and benign tumors[21]. Due to the increased cellular density and

corresponding low diffusion rate found the cancerous tissues[22], these areas of the prostate gland are shown to have lower ADC values when prostate cancer is present.

Current PCa Imaging Methodology and Methods

A review of current literature shows two preferred computer aided MRI image analysis methods for prostate cancer detection and classification: Convolutional Neural Networks (CNN)[8,23,24,25], and texture feature extraction and classification[3,9,26,27,28]. In addition to using T2W MRI to aid in the detection and classification of prostate tumors, several previous studies[10,21,22,29,30] show that combining DWI MRI with T2W MRI will produce better classification accuracy than can be obtained using T2W images alone.

Since this project will focus on texture feature extraction and classification methods, it is that literature which has been reviewed. Table 1 shows a summary of the top texture feature extraction methods[3,26,28], dimension reduction methods[3,26,31], and classification methods[10] found in current literature.

	Popular Methods
Feature Extraction	Grey Level Co-occurrence Matrix (GLCM), Wavelets, First order statistics (Mean, median, standard deviation)
Dimension Reduction	Minimum-Redundancy–Maximum-Relevancy (mRMR), Lasso and Ridge Regression
Classification	Support Vector Machine (SVM), Random Forests

Table 1. Top feature extraction, dimensionality reduction, and classification methods obtained from current literature.

Project Definition

Evaluate the feasibility of machine learning methods based on feature extraction and classification for differentiating benign and malignant prostate cancer tumors in mp-MRI with the intent of building an algorithm that best reduces over-biopsy and enables in-silico biopsy. The model will use regions of interest (ROI) in 2D multiparametric magnetic resonance images (mp-MRI) selected by the radiologist as possible prostate cancer tumors. This project will evaluate various extracted texture features, dimensionality reduction, and machine learning classification methods to determine the computer-aided analysis model, or models, that provide the greatest classification accuracy for malignant and benign prostate cancer images.

Chapter 2: Materials and Methods

Image dataset

Both T2W and DWI sets have been obtained from a cohort of 81 patients and have been correlated with biopsy results for both malignant and benign tumors. In this dataset the classification of malignant is defined by a Gleason score of ≥ 6 and a benign classification is defined as not having a Gleason score. All of the patients in this study had previously undergone TRUS biopsy that yielded indeterminate results and thus were re-screened using the MRI in-bore guided biopsy method. It was from this second biopsy that the Gleason scores and malignant/benign classification were obtained.

Of the 81 patients currently in the dataset, 3 were omitted from the dataset due to inconclusive biopsy results (Atypical small acinar proliferation), yielding 78 patients for this study. It should be noted that only images that were deemed as possibly cancerous by a radiologist have been included in this dataset, and as a consequence, all patients do not all have the same number of images. Table 2 shows the distribution of patients and their associated images for the benign and malignant classes, patient image set size range, and the distribution of images for the transition zone (TZ) and the peripheral zone (PZ) of the prostate gland.

	Patients	**Images**	**Patient image set size range**	**PZ Images**	**TZ Images**
Benign	23	64	1 - 6	21	43
Malignant	55	155	1 - 8	50	105
Total	78	219		71	148

Table 2. Distribution of benign and malignant patients and images.

As one can see, the dataset is unbalanced having more malignant patients and images than benign. The unbalanced distribution of the dataset will cause a bias in prediction towards the malignant class at the expense of misclassifying the benign class. In an attempt to balance the dataset between malignant and benign images, an oversampling technique will be used where all 64 benign images, and associated patients, will be reused during the training and test process. This method will essentially increase the benign image set to 128 and leave the malignant image set at 155, while allowing the study to include all 78 patients and their associated images. Although the oversampling method better balances the dataset, an imbalance of 27 benign images still exists. The exact details regarding the implementation of this dataset balancing method are outlined in the *model evaluation* section later in the document.

Figure 2 serves to highlight differences and similarities between T2W and DWI ADCmap images for both malignant and benign cases. As can be seen the ADCmap images show a better contrast between the ROI and healthy tissue than do the T2W images

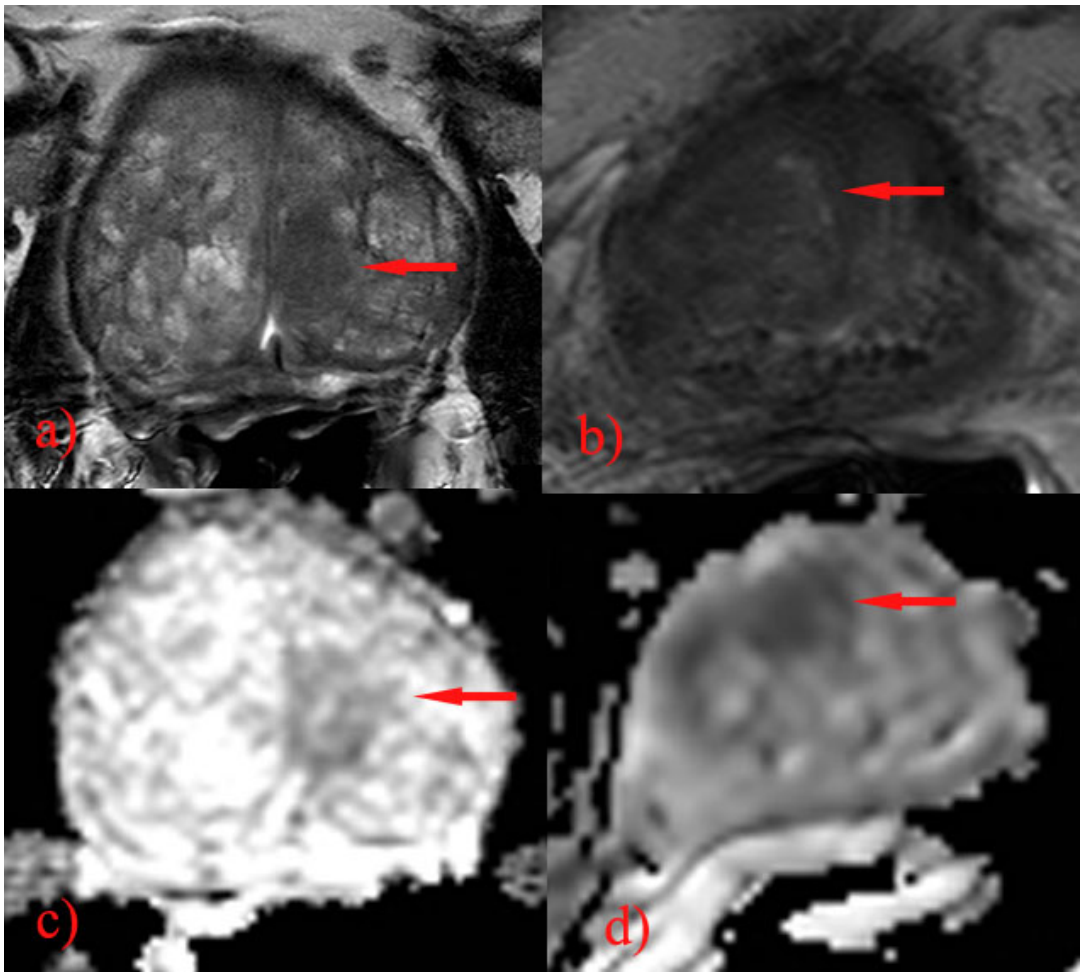

Figure 2. Malignant and benign image comparisons. The ROI for each image is indicated by the red arrow. (a) Benign T2W image, (b) malignant T2W, (c) benign ADCmap with, and (d) malignant ADCmap.

To aid in image analysis efforts, the regions of interest (ROI) have been outlined by radiologists for the T2W and ADC images, and will be used by the computer image analysis models for classification of images. Figure 3 shows examples of both types of mp-MR images obtained for this project.

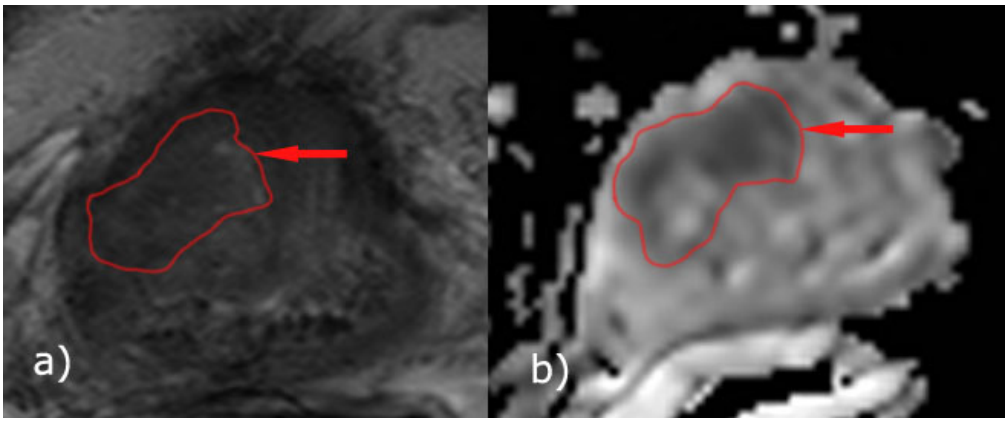

Figure 3. T2W and DWI ADC ROI Images. (a) shows a T2W image with the ROI outlined in red, indicated by the red arrow, (b) shows a DWI ADC map with region of interest indicated by the red arrow.

Image analysis workflow

The image analysis workflow in Figure 4 applies an image pre-processing step to de-noise and generate extracted ROI images from both the T2WI and ADCmaps.

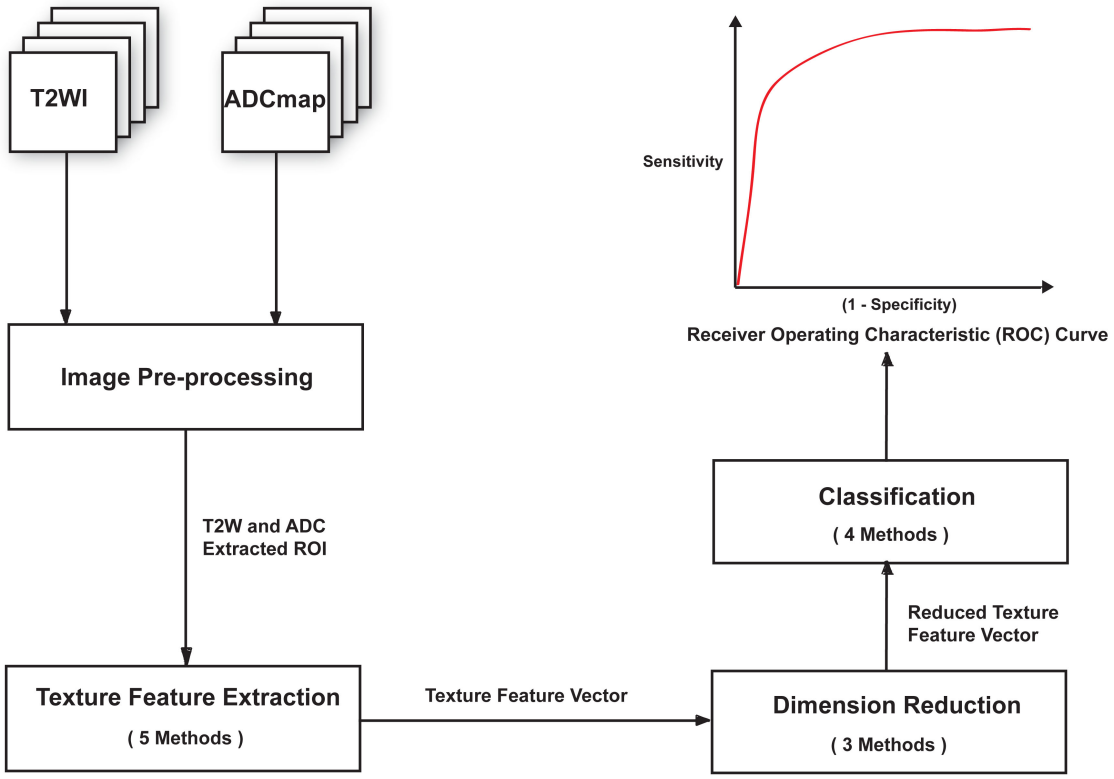

Figure 4. Image analysis algorithm

Both original and de-noised versions of the T2WI and ADCmaps are used to create the ROI images. Once the ROI images have been created, five feature extraction

methods, described in the next section, are used to create texture feature vectors from these images. The appropriate malignant and benign class is assigned to each feature vector for use in later supervised learning steps. In order to identify features that best describe each class, three dimensionality reduction and feature selection methods have been used: maximum relevance minimum redundancy (mRMR), autoencoder, and Elastic Net, are applied to the full texture feature vectors to produce reduced texture feature vectors associated with each class. The reduced texture feature vectors are then used by three supervised machine learning classification techniques: Support Vector Machine (SVM), Random Forest, and Linear Regression, to classify the images, and produce a Receiver Operating Characteristic (ROC) curve. The ROC curve will be used to analyze the accuracy of the various methods in order to obtain the best image analysis model.

In determining the best image analysis model, cross combinations of the 5 texture feature extraction methods, 3 dimension reduction methods, and 3 classification methods will be used. In addition, the T2W and ADC ROI images will be used separately and together, as inputs to each cross-combination model, to determine which image modality best predicts the malignant and benign classes. The final goal is not only to develop the best prediction image analysis algorithm, but also to gain an understanding of the MRI modality type and image features that can best predict malignant and benign prostate cancer in the given dataset.

Image Pre-processing

In this study the image pre-processing step contains two elements: the removal of any unwanted noise "de-noising" from the original images, and extraction of the ROI portion from the complete MR image.

In any image, unwanted noise is added to the image during the acquisition process and/or during subsequent processing steps. It is because of this unwanted noise, that image de-noising techniques are employed in an attempt to restore the original image integrity for more accurate analysis. Several reports list noise in MRI images follows a Rician distribution[32,33] while another report indicates MRI noise consists mainly of Salt and Pepper, Speckle, Gaussian and Poisson noise[34]. Many image de-noising techniques exist, from which three are used in this study: the median filter, variational autoencoder, and the convolutional autoencoder. The order-specific median filter is perhaps the simplest and most used de-noising filter in image processing. The median filter is popular for its ability to remove or reduce random noise, and is specifically effective in removing salt and paper (impulse) noise while adding less blurring than similar sized smoothing filters[35]. The median filter works by replacing the intensity value of a given pixel with the median values of pixels in a given local neighborhood around that pixel.

Autoencoders form the basis for two of the noise reduction techniques used in this study. Since an autoencoder is also used as a dimensionality reduction technique described later in this paper, it is appropriate to discuss the basic autoencoder function at this time. An autoencoder is a fully connected neural network (NN), with the goal of the output reproducing the values present at the input of the network. Figure 5 shows a three layer autoencoder consisting of an

input and output layer of *m* neurons, and a hidden layer of *n* neurons, where n < m. The autoencoder is composed of two parts: an encoder and a decoder. The hidden layer shown in Figure 5, also known as the latent layer, is an encoded version of the autoencoder's input variables.

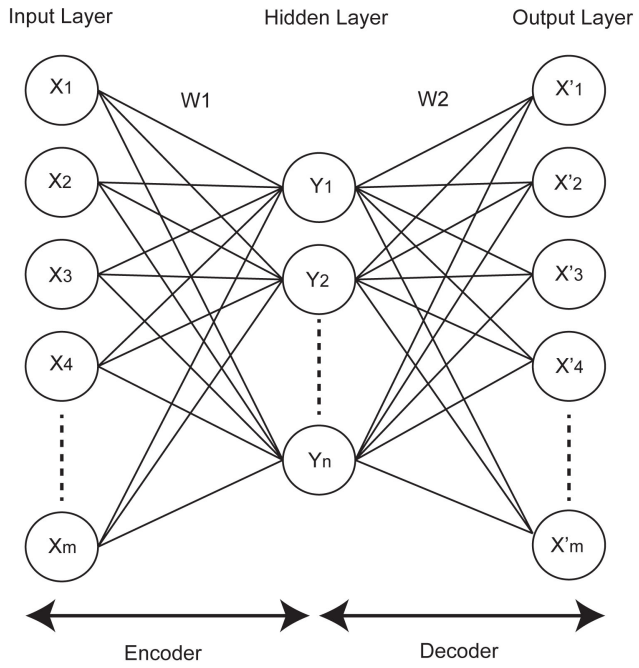

Figure 5. Simple autoencoder configuration.

From the base autoencoder concept, two types of autoencoder are shown[36,37] for image de-noising: the De-noising Autoencoder (DAE), and the Variational Autoencoder (VAE). The DAE follows the architecture of the basic autoencoder shown in figure 5. In the encoder portion of the DAE, the input is converted into the latent encoded vector, which contains a single value for each encoded dimension. One issue with the encoding method of the DAE as a generative model is that its encoded latent layer may not be continuous. The discontinuous nature of the DAE's latent layer will tend to cluster images together, making it useful for replicating the same images, but does not allow easy interpolation of images that differ from those used to train the model.

In contrast to the DAE, the VAE encodes an input in a probabilistic manner. This probabilistic encoding provides a continuous latent space, and describes each latent variable as a probability distribution and not a single value as does the DAE. The decoder portion of the VAE randomly samples from the latent distributions to reconstruct the input at the output of the encoder. In order to achieve a continuous latent space, it is assumed the prior follows a normal distribution, and the VAE splits up the latent vector into two vectors, one of the input means, and the other of the input standard deviations as shown in Figure 6. The VAE decoder samples and generates the sample layer from the latent distribution layers, mean and standard

deviation, created from each input variable. This sampled layer is the input to the decoder, which attempts to reconstruct the input at the output layer of the network.

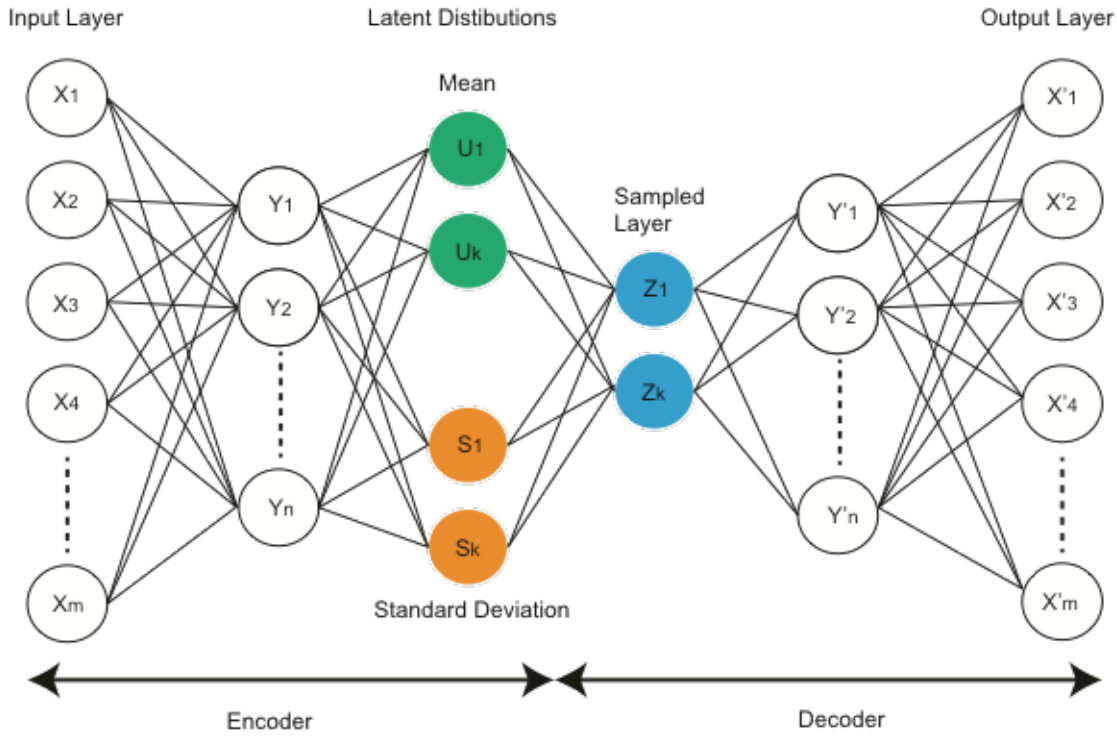

Figure 6. Variational Autoencoder. Latent distributions showing the mean and standard deviation latent layers and sampled input layer to the decoder.

Because the VAE encodes a continuous latent space that allows for interpolation between encodings, it is desired that these encodings be as close as possible while still being distinct, allowing for smooth interpolation. To make sure that the encodings follow this desire, the Kullback–Leibler divergence (KL) term is added to the loss function:

$$L(\theta,\phi;x_i) = -D_{KL}(q_\phi(z \mid x_i) \| p_\theta(z) + E_{q_\phi(z|x_i)}[\log p_\theta(x_i \mid z)] \quad (2)$$

where the first term is the KL divergence containing the continuous latent variable $q(z \mid x_i)$, and the second term is expected reconstruction error[37] containing the reconstructed image $p_\theta(x_i \mid z)$. In this study, both binary cross-entropy and mean squared error will be used as the reconstruction error function to determine which, if either, yields the best network performance.

In practice, due to random sampling of the latent distributions used to calculate the sampled layer, model training using backpropagation becomes very difficult[37,38]. We can still use backpropagation however by employing a method knows as the "reparameterization trick"[37,38]. The "reparameterization trick" takes a random sample ε from a unit Gaussian distribution; the sample ε is then shifted by the mean

of the latent distribution, and scales it by the latent distributions standard deviation. This yields a new sampled distribution Z in the form:

$$z = \mu + \sigma \otimes \varepsilon \tag{3}$$

and the new loss function can be written as:

$$L(\theta,\phi;x_i) \approx 0.5\sum_{j=1}^{J}(1+\log(\sigma_j^2) - \mu_j^2 - \sigma_j^2) + \frac{1}{L}\sum_{l=1}^{L}\log(p_\theta(x_i \mid z_{i,l})) \tag{4}$$

where $z_{i,j} = \mu_i + \sigma_i \otimes \varepsilon_l$

For the VAE implementation in this project, the input images are cropped to a standard size of 200x200 pixels centered on the ROI. The cropped images are then flattened into a vector of 40,000 pixel intensity values that will be the input to the VAE and also used at the output of the VAE during training. Hidden layers Y and Y' consist of 2000 nodes fully connected to the input and output layers. Finally, the latent layers mean and standard deviation, along with the sampled layer, contain 200 nodes each.

In addition to using de-noising autoencoders such as the VAE, that use one-dimensional input vectors constructed from an image, a two-dimensional input Convolutional Autoencoder (CAE) can be used for image denoising[39], and dimensionality reduction. The CAE is based on the conventional autoencoder model containing both encoding and decoding layers. Unlike the autoencoders discussed thus far, which have fully connected layers, the CAE is based on Convolutional Neural Network (CNN) architecture. The CNN has become a mainstay in computer vision and has also shown to be valuable in prostate MRI analysis and classification[8,23,24,25]. A CNN is like a conventional neural net (NN) in that the CNN contains hidden layers of neurons that have weights and bias that can be configured during the learning process. The CNN also shares similar computations to the NN such as sum of products, added bias, and results passed through activation functions, where the activation value becomes the input to the next neuron. However, unlike conventional neural networks that use a vector as an input, the CNN uses 2-D arrays (images) thus making them ideal for analyzing and classifying images. One advantage of using a CNN over a NN for image analysis is that a NN does not scale well to full images due to their fully connected layer architecture. This fully connected architecture rapidly increases the number of learnable weights as the image size increases and will become computationally unmanageable and possibly lead to overfitting of the net. Because the CNN assumes the inputs are images the architecture can be modified to be more efficient for image specific tasks utilizing a 3D connection of neurons. Figure 7 shows a LeNet architecture of a CNN as an example of the general elements (layers) contained in a CNN.

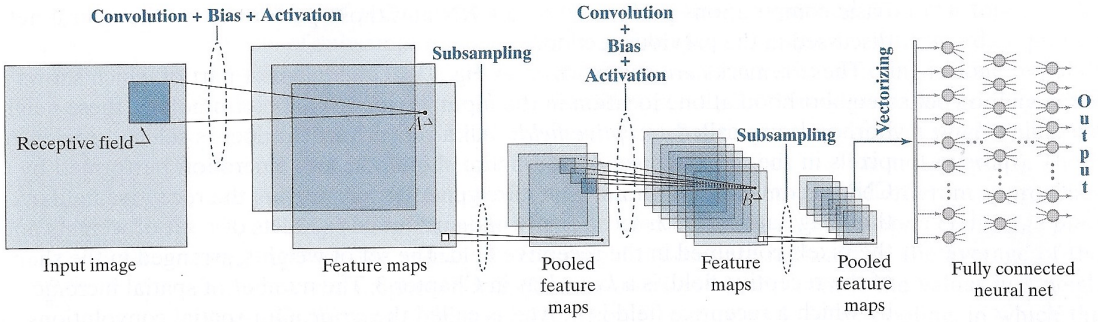

Figure 7. Example of a CNN[35].

The input stage of the CNN employs neighborhood processing of a receptive field in the input image. The computation of the receptive field employs sum of products and set of kernel weights and is commonly known as the convolution layer. Each convolution produces a single value to which a bias is added and passed to an activation function. There are several choices of activation functions including sigmoid, tanh, and rectified linear unit (ReLU). This layer is commonly known as the activation layer and can be thought of as the output of the neuron. A 2D array, called a feature map, is created from the output of the activation layer. The process after convolution and activation is subsampling (or pooling), and is used to reduce the size of the feature maps creating a set of pooled feature maps (one pooled feature map for each feature map). In addition to subsampling pooling helps with translational and rotational invariance. There are several methods of reducing the feature maps including averaging of values in a field or the maximum of value in a field. Commonly a 2x2 field, or pooling neighborhood, is used in the subsampling layer. The convolution, activation, and pooling layers can be repeated a number of times to tune the CNN for optimal performance. In Figure 7, we observe two sets of convolution, activation, and pooling layers with the inputs to the 2nd convolution layer being the pooled feature maps. In the final output stage of the CNN, the pooled feature maps are vectorized to create a single input vector used as the input of a fully connected neural net (FC layer).

Unlike the general example of the CNN shown in Figure 7, the CAE used in this project, shown in Figure 8, does not have a final FC layer since the CNN is being used as an autoencoder. Like the VAE, the input images are cropped to a standard size of 200x200 pixels centered on the ROI. These images are used at both the input and output layers during training of the CAE. The feature maps and pooled feature maps have a depth of 32 and the size are scaled down twice in the encoder, and scaled up twice in the decoder. All activation functions are ReLU with the exception of the output convolution layer where a sigmoid function is used.

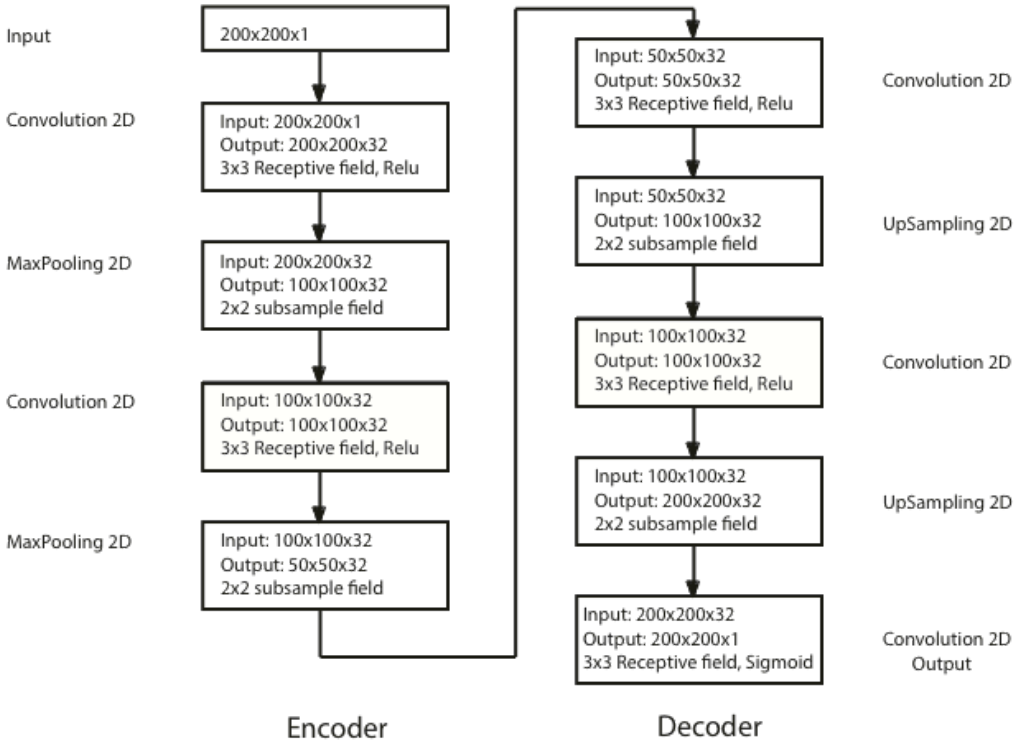

Figure 8. CAE as implemented for this project.

The second step in the image pre-processing phase is ROI extraction, using either the original or de-noised images. The original image dataset contains both modes of the MR images, and ROI masks obtained from a radiologist. These masks are images where the radiologist outlined ROI area with white (255 grey scale) and the remaining image is black (0 grey scale), and are aligned with the MR images. By normalizing the ROI mask, and then multiplying the mask with the MR image, a new image with just the ROI visible. To reduce image storage space and compute time, the new ROI image is cropped to the size of the visible ROI, to produce the final ROI image as shown in Figure 9. It is this final cropped ROI image that will be used for feature extraction.

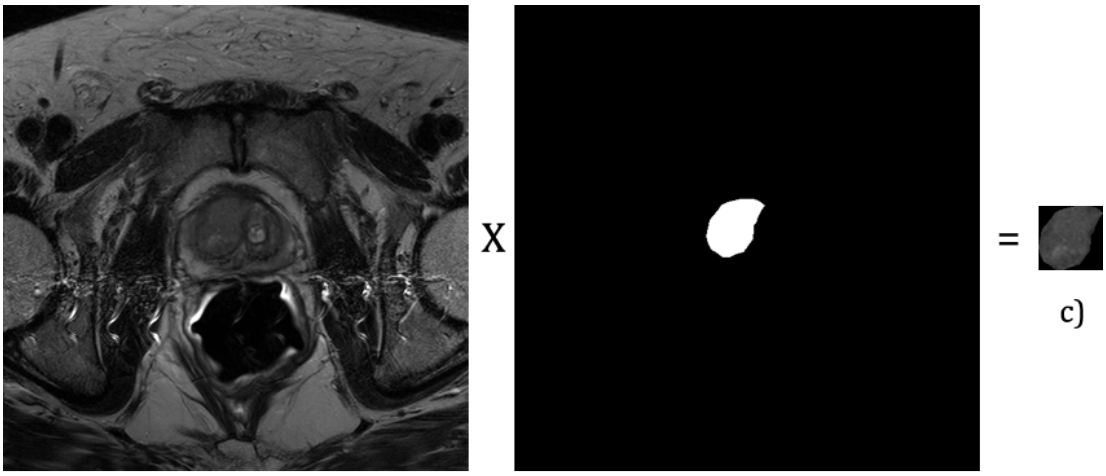

Figure 9. Extraction of a ROI image. (a) Original or de-noised MR image, (b) ROI mask, (c) Resulting cropped final ROI image

Texture Feature Extraction

The term texture is generally accepted as a way to characterize a region of an object. There are many ways to characterize a texture, with some of the most commonly used methods being periodicity, intensity, and heterogeneity. In this study, texture feature extraction of the image ROI areas will be accomplished using three statistical matrix methods: grey-level co-occurrence matrix (GLCM), grey-level size zone matrix (GLSZM), and grey-level run length matrix (GLRLM). In and effort to make the matrices less sensitive to noise, one approach is to quantize the intensities (grey-levels) into a smaller number of groups. Before feature extraction, the pixel intensity values will be binned in sizes of: 1 (original image), 8, 16, and 32. As an example of bin size 8, intensity values from 1-7 are set to 1, intensity values of 8-15 are set to 2 and so on. In addition, Multiresolution Fractal Dimension analysis and Local Binary Pattern (LBP), will also be used. Thibault et. al.[40,41] discuss the use of these statistical matrices in cell nuclei classification and the analysis of MR images in breast cancer therapy response[42]. This project will evaluate these methods for the possibility of enhancing the analysis and classification of prostate cancer mp-MR images.

The GLCM is a popular and widely used statistical method of texture feature extraction based on the spatial relationship of pixels in an image by calculating the second order statistics of image texture. The GLCM is an NxN matrix where N in the total number of grey level intensity values for a given image. The $(i,j)^{th}$ entry in the matrix represents the total number of times a pixel with a grey level of i is separated from another pixel with a grey level value of j separated by a distance k for a given displacement vector $\vec{d} = (d_x, d_y)$. This calculation is shown in equation 5:

$$M_d(i,j) = card\left\{\begin{matrix}((r,s),(r+d_x,s+d_y))/ \\ I(r,s)=i, I(r+d_x,s+d_y)=j\end{matrix}\right\} \quad (5)$$

Figure 10 shows the adjacency directions (0°, 45°, 90°, 135°) that make up the four matrices from which the Haralick features are calculated.

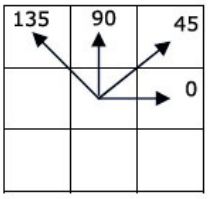

Figure 10. The four directions of adjacency.

Figure 11 shows and example of the construction of a GLCM using a pixel pair relationship of d=1, angle = 0°. It can be seen that the pair (1,1) occurs only once in image *f* where the pair (6,2) occurs 3 times.

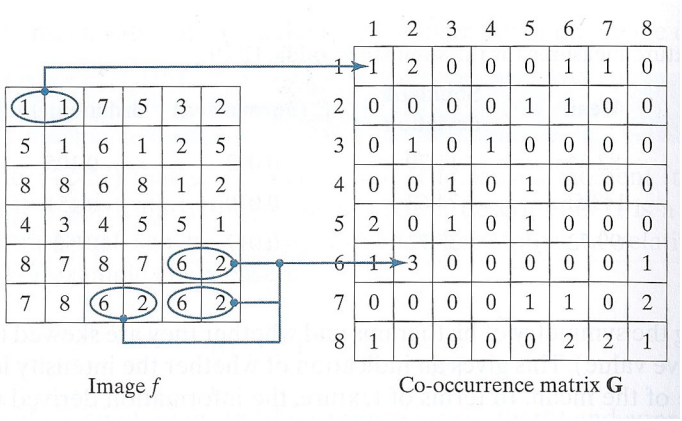

Figure 11. Co-occurrence matrix construction example for an image consisting of 8 grey levels[35].

In the example shown in figure 10 there are only 8 grey levels making the co-occurrence matrix relatively small, however for images with larger numbers of grey levels, e.g. 256 or even 65535, the co-occurrence matrix can become quite large and computationally expensive. There are a number of useful statistical descriptors that can be used to characterize a GLCM. In this study, 13 Haralick features[43,44] will be extracted from the ROI images. Note that intensity 0 is considered background around the ROI and is not used in any feature extraction calculations by the software. Once the image intensity values have been binned, the features are extracted and feature vectors generated. The feature vectors are generated in two ways: a composite of the 13 Haralick features generated in each of the four

directions, yielding a vector of length 52, and a rotationally invariant feature set averaging all four directions, yielding a vector of length 13.

The LBP texture analysis method is known for its grey scale invariance, robustness to noise, and characterization power. The general LBP method uses a pixel neighborhood, usually 3x3, the pixel intensities in the neighborhood are thresholded using the center pixel value, a pattern is computed, and the pattern value assigned to the center pixel. Figure 12 shows the steps used to calculate the neighborhood LBP score.

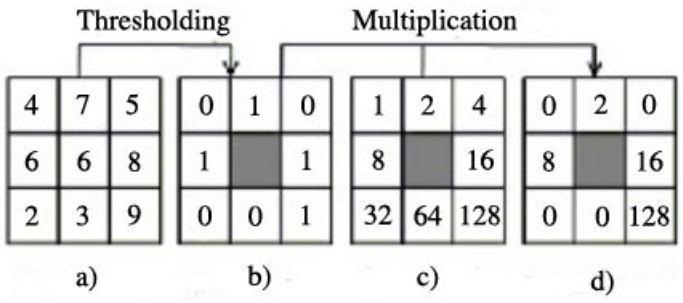

Figure 12. Steps used to calculate the neighborhood LBP score. (a) Original image 3x3 neighborhood, (b) thresholding derived binary values, (c) weight matrix, (d) resulting matrix used to calculate LBP score[45].

Figure 12a shows the image neighborhood used to calculate the LBP score for the center pixel. The center pixel is also used as the threshold value (in this case 6) to create the binary values of the neighboring pixels shown in Figure 12b. If a pixel is below the center pixel threshold value then it is assigned 0, or 1 otherwise:

$$G(x, y) = \begin{cases} 0 & I(x,y) < I(0,0) \\ 1 & \text{Otherwise} \end{cases} \tag{6}$$

Where G(x,y) is the assigned binary value, I(0,0) is the center pixel intensity value (threshold) and I(x,y) is the intensity value of the neighboring pixels. The resulting threshold binary values undergo and element-wise multiplication with weighted values (figure 12c) to produce a result matrix shown in figure 12d. The LBP score is determined by summing the resulting values in figure 12d, in this example the LBP score is: 2+8+16+128 = 154. All resulting LBP scores are used to build a histogram, which represents the structural texture characteristics of the image.

For a 3x3 neighborhood, 8 pixels are used to calculate the LBP, known as the LBP^8 operator[46]. The LBP^8 operator generates a total of 256 unique binary patterns from the 8 pixel values surrounding the center pixel. If the image is rotated, a different pattern contained in the 256 pattern set will be selected. To remove this rotational effect, 36 unique rotational invariant patterns can be identified using the definition:

$$LBP_8^{ri36} = \min\left\{ROR(LNP_8, i) \mid i = 0,1,\ldots,7\right\} \tag{7}$$

where *ROR(x,i)* is a right clockwise pixel rotation that yields the maximal number of most significant bits as 0, or a right shift of the 8-bit binary pattern x, i number of times. LBP_8^{ri36} can be considered a feature detector as it quantifies the occurrence of these 36 patterns to various features contained in the image. This study will use a histogram of the 36 rotationally invariant patterns to form the LBP feature vector for each MRI modality (T2W and ADC). Two histograms will be created, the standard LBP normalized histogram that shows the probabilities of each pattern extracted from an image, and a non-normalized histogram that contains pattern counts for the 36 patterns calculated from the image. A second modification to the standard LBP methodology is to analyze the effects of grey-level intensity binning, as used in the GLCM, GLRLM, and GLSZM, compared to the standard method of analyzing each grey level individually, which in this case is 256 grey-levels.

The GLRLM is another widely used statistical matrix technique for texture characterization[47]. In the GLRLM a run is defined as the consecutive pixels with the same value in a given direction. As with the GLCM, the directions used to calculate run length are (0^o, 45^o, 90^o, 135^o). Since run calculations are symmetric, it is not necessary to use the directions (180^o, 225^o, 270^o, 315^o)[25]. Each element of the GLRLM contains the number of runs for a given grey-level and run length. An example of the calculation of a GLRLM using an image with 4 grey levels is shown in Figure 13.

Image					Gray Level (i)	Run Length (j)			
						1	2	3	4
1	2	3	4	=>	1	4	0	0	0
1	3	4	4		2	1	0	1	0
3	2	2	2		3	3	0	0	0
4	1	4	1		4	3	1	0	0

Figure 13. Example of the calculation of a GLRLM using an image with 4 grey levels[25].

It can also be seen in Figure 13 that the height of the matrix is dependent on grey-levels, and the number of columns in the matrix will be dynamic, dependent on the size of the longest run length. After the GLRLM has been constructed statistical measures of moments from -2 to 2 for the extracted texture features can be calculated. In this study, 11 joint statistics will be calculated as proposed in[43,44] and are shown in Table 3.

Feature	Formula
Short Run Emphasis (SRE)	$\frac{1}{n}\sum_{i,j}\frac{p(i,j)}{j^2}$
Long Run Emphasis (LRE)	$\frac{1}{n}\sum_{i,j} j^2 p(i,j)$
Grey Level Non-uniformity (GLN)	$\frac{1}{n}\sum_{i}(\sum_{j} p(i,j))^2$
Run Length Non-uniformity (RLN)	$\frac{1}{n}\sum_{j}(\sum_{i} p(i,j))^2$
Run percentage (RP)	$\sum_{i,j}\frac{n}{p(i,j)j}$
Low Grey Level Run Emphasis (LGRE)	$\frac{1}{n}\sum_{i,j}\frac{P(i,j)}{i^2}$
High Grey Level Run Emphasis (HGRE)	$\frac{1}{n}\sum_{i,j} i^2(i,j)$
Short Run Low Grey Level Emphasis (SRLGE)	$\frac{1}{n_r}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{p(i,j)}{i^2.j^2}$
Short Run High Grey Level Emphasis (SRHGE)	$\frac{1}{n_r}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{p(i,j)i^2}{j^2}$
Long Run Low Grey Level Emphasis (LRLGE)	$\frac{1}{n_r}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{p(i,j).j^2}{i^2}$
Long Run High Grey Level Emphasis (LRHGE)	$\frac{1}{n_r}\sum_{i=1}^{M}\sum_{j=1}^{N} p(i,j).j^2 i^2$

Table 3. The 11 statistical calculations used for the GLRLM.

As with GLCM, images will be binned in sizes of 1 (original image), 8, 16, and 32 image intensity groups. After binning, features will be extracted to form feature vectors of length 44 when using the four directions (0^o, 45^o, 90^o, 135^o), and a feature vector of length 11 will be calculated for the rotation invariant averaging of these four directions.

As a complement to the GLCM and GLRLM, Thibault et. al.[41] developed the GLSZM to characterize the homogeneity of an image texture by analyzing zones by their size and grey-level intensity. A homogeneous texture will contain large areas of the same grey-level intensities called flat zones. The more homogeneous the texture the wider and flatter the matrix. An example of the computation of a GLSZM is shown in Figure 14.

Image

1	2	3	4
1	3	4	4
3	2	2	2
4	1	4	1

=>

Gray Level (i)	Size Zone (j)			
	1	2	3	4
1	2	1	0	0
2	1	0	1	0
3	0	0	1	0
4	2	0	1	0

Figure 14. Calculation of a GLSZM from and image containing 4 grey-levels[40].

The size of the zone is calculated from pixels of the same grey-level connecting in any direction. In this example note that grey-levels of 2, 3, and 4 all have zone sizes of 3 but are connected in different directions; grey-level 3 in the diagonal direction, grey-level 2 in the horizontal direction, and grey-level 4 in both the vertical and horizontal directions. Also, like the GLRLM, the GLSZM contains a fixed number of rows determined by the number of chosen grey-levels and is dynamic with respect to the number of columns which are determined by the size zones found in the image. Another aspect of the GLSZM is that it is invariant to image translation and rotation whereas the GLCM and GLRLM are dependent on the offset and orientation used in their calculations. The same 11 joint statistics used in the calculation of the GLRLM will be used to calculate the GLSZM and are shown in Table 3. For clarity, "Run" is replaced with "Area" in the feature descriptions.

There are many ways to extract fractal-based texture features from and image. In this study the multiresolution fractal dimension texture analysis technique[45] will be investigated for texture feature extraction. The multiresolution fractal dimension technique for texture feature extraction is comprised of two steps: wavelet image decomposition and fractal dimension analysis of the decomposed image.

Most tissue texture is shown to be heterogeneous, and as such, has more complicated texture features than does a homogeneous texture. It is because of this texture heterogeneity that single frequency wavelet decomposition may not capture all of the texture features and thus a multiresolution approach is required. It is proposed that a set of 4 different frequency Gabor wavelets be used to decompose the image texture into sub-bands in order to capture multiple frequency channels of the texture.

Fractal dimension will be calculated from the Gabore wavelet decompositions of the image using the Differential Box Counting (DBC) method[50,51]. The fractal dimension calculation in the DBC method is given by equation 8:

$$FD = \lim_{r \to 0} \frac{\log Nr}{\log(1/r)} \qquad (8)$$

In this method the image is divided into grids of size SxS from the original MxM image matrix. The divisions are in increments of 1/2n for n {1,2,4,...} until a 4x4 minimum grid size is achieved. Quantized intensity level boxes are calculated and placed on the image in what could be considered the z-dimension covering each SxS grid. The granularity H (height in the z-dimension) of the grey level boxes is computed using: H = G (S/M), where G is the total number of grey levels present in

the MxM image. This yields SxSxH 3-dimensional boxes covering the MxM image. To compute the number of grey level boxes covering each SxS grid, we identify the lowest grey level in the grid and assign the box number containing that grey level to *k*, and find the highest grey level in the grid and assign the box number containing that level to *l*. Thus we can compute the thickness of the boxes covering a specific SxS grid (i,j) as:

$$n_r(i,j) = l - k + 1 \tag{9}$$

Over the entire MxM image, $n_r(i,j)$ can be summed in equation 10:

$$N_r = \sum_{1<i<M,1<j<M/S} n_r(i,j) \tag{10}$$

The fractal dimension FD can be estimated from the least squares linear fit of a plot of log(Nr) vs. log(1/r) where r = S/M.

Once the fractal dimensions of the 4 wavelet decompositions are calculated, the sub-band decomposition with the largest fractal dimension is then further decomposed using the 4 wavelet frequencies and these 4 decompositions are then used to calculate fractal dimensions. Figure 15 shows an example of the tree structure of the wavelet/fractal steps.

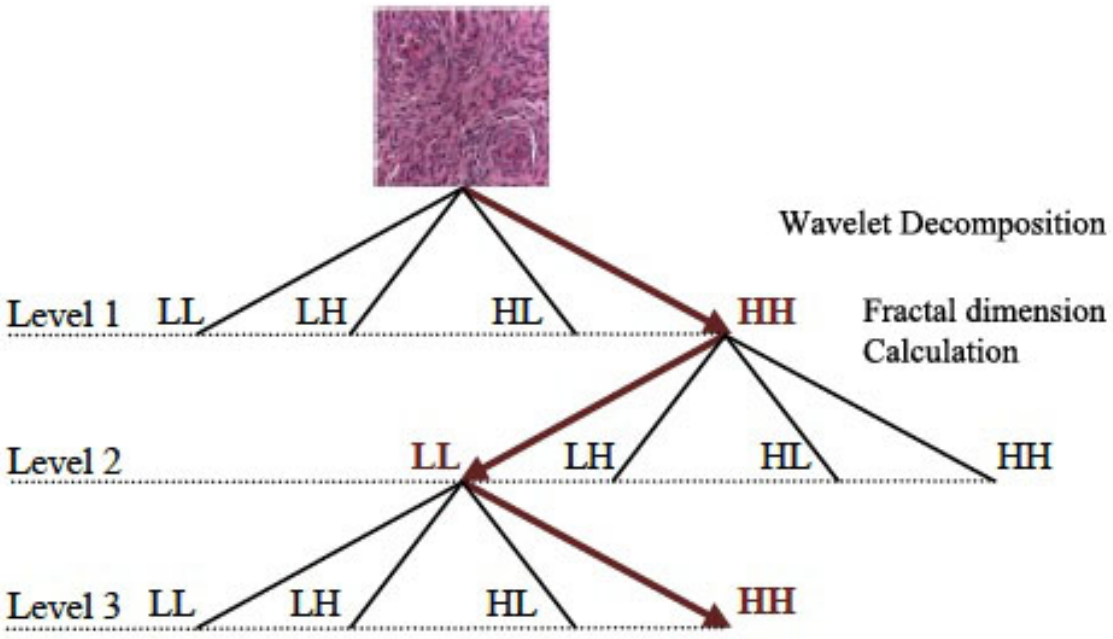

Figure 15. Wavelet quad-tree consisting of three layers[49]. Wavelet decomposition is performed followed by fractal dimension calculation to determine the tree branch point.

In some cases, an in-between difference threshold is set to determine the termination of the tree when all fractal dimensions are within this threshold. Here we wish to keep the number of fractal dimensions constant in our texture feature vector so a tree depth of 4 will be set for all images, resulting in 16 fractal dimension texture features.

Table 4 shows a summary of the number of features and the number of feature vectors calculated from the five feature extraction methods used in this study.

Feature Extraction Method	**Number of Extracted Features**	**Number of Feature Vectors**
GLCM	13 per adjacency direction, for a total of 52 for all four directions. 13: average of four directions	8 total: 4 adjacency directions and average for each of the 4 intensity bin sizes
GLRLM	11 per adjacency direction, for a total of 44 for all four directions. 11: average of four directions	8 total: 4 adjacency directions and average for each of the 4 intensity bin sizes
GLSZM	11	4 total: 1 for each intensity bin size
LBP	36	8 total: 4 normalized for each of the intensity bin sizes, and 4 non-normalized histograms for the 4 intensity bin sizes
MRFD	4 per adjacency direction for a total of 16 for all four directions. 4: average of four directions	2 total: for adjacency directions and average of the 4 directions.

Table 4. Numbers of extracted features and generated feature vectors for each feature extraction method.

A total of 30 feature vectors are created for each of the T2W and ADC image modalities. These 30 feature vectors from each image modality form the basis for calculating the multiparametric feature vectors. The following steps were used in creating the multiparametric feature vectors from the single mode T2W and ADC feature vectors:

1) The 30 feature vectors from the T2W and ADC images were used with all models to determine a rank of features by AUC scores for each image modality.
2) From the ranked lists of AUC scores from the 30 T2W and ADC features, the top 5 features from each image modality were selected, a vector **T5** = [t1, t2, t3, t4, t5] for T2W and a vector **A5** = [a1, a2, a3, a4, a5] for ADC.
3) The vectors **T5** and **A5** are combined, $\mathbf{A5} \otimes \mathbf{B5} = \mathbf{A5^T}\ \mathbf{T5}$ = a 5 x 5 matrix containing the 25 top T2W and ADC combinations. The values from this matrix from a vector of 25 features **M1** = [f1, f2, …, f25].

4) A second vector, **M2**, of 30 features was made from the adding the same feature types from the T2W and ADC feature sets. Let **T24** = [t1, t2, …., t24], **A24** = [a1, a2, …, a24] and **M2** = **T24** + **A24** = [t1a1, t2a2, …… t24a24]. As an example, t1 is GLCM Bin size 1 for T2W images, and a1 is GLCM Bin size 1 from ADC images.
5) The multiparametric features in M1 and M2 were used with all models to determine a rank of features by AUC scores. The top 5 multiparametric features, m1 – m5 were combined to form a set of combined multiparametric features: {m1m2, m1m3, m1m4, m1m5, m2m3, m2m4, m2m5, m3m4, m3m5, m4m5}.
6) The set of combined multiparametric features determined in step 5 were then used with all models to determine their AUC scores.

In total, 65 multiparametric features were created for model evaluation. The rational for developing multiparametric features in this manner was to determine which feature extraction method, or combination of methods, performed the best for each model evaluated.

Dimensionality Reduction

In general, an important step in machine learning is dimensionality reduction, or feature selection, that aims to find the subset that best predicts. Dimensionality reduction is important for several reasons: highly dimensional data sets tend to be very sparse and are prone to overfitting the model, decrease computation time, and some methods improve interpretation of the data for example to identify patterns or clusters. The exception to the features interpretation after dimensionality reduction is the autoencoder, which make the data impossible to interpret. In this project three dimensionality reductions methods are proposed: Minimum-Redundancy–Maximum-Relevancy (mRMR), Autoencoder, and ElasticNet.

Minimum-Redundancy–Maximum-Relevancy (mRMR) is a dimensionality reduction algorithm that has become popular in texture feature extraction and classification[3,26,31]. In order to select the most relevant subset of features one must measure the dependence between random variables. In information theory a widely used method to determine dependence is the concept or mutual information. In the mRMR approach, mutual information can be used to measure both the relevance and redundancy of random predictor variables[52]. The Maximum-Relevancy portion of the algorithm determines the feature subsets with the highest correlation to a prediction or classification value, while the Minimum-Redundancy portion of mRMR removes redundancy from these subsets by determining the largest correlation between features. These two operations are performed in parallel to find a balanced tradeoff between relevance and redundancy in order to determine the optimal minimum reduced feature set. In the implementation of mRMR used in this study, two thresholds can be set to determine both the relevance range w.r.t. the feature variable with the largest correlation to the classifier, and a threshold that specifies the redundancy, or correlation, to the other feature predictors. These thresholds are

set using several feature vector types to determine the best performance of both the SVM and Random forest classifiers.

As mentioned in the image pre-processing section, the autoencoder can be used for a variety of purposes such as image de-noising and dimensionality reduction. As the dimensionality reduction method used in this study, the autoencoder reduces the input vector X to a smaller latent vector Z as shown in Figure 16. The latent vector Z will be used as the reduced predictor feature set used by the SVM and Random Forrest classification methods.

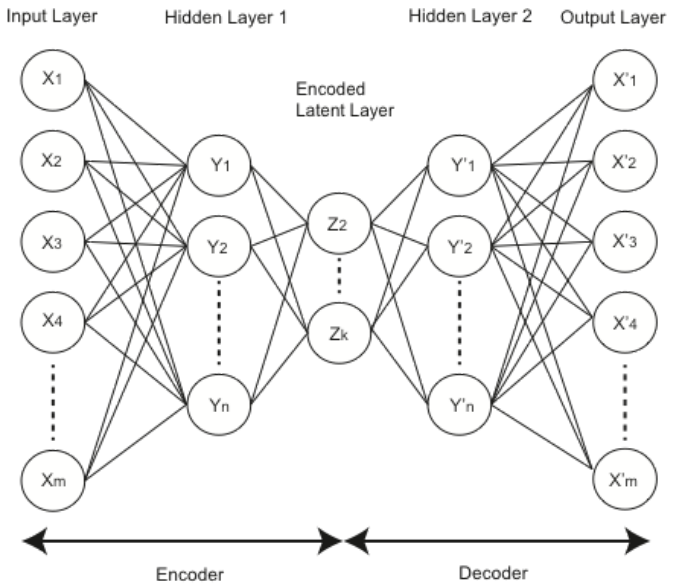

Figure 16. Dimensionality reduction autoencoder.

During training, the input and output are set to the training set of feature vectors at X and X'. A constraint of m > n > k is placed on the model where m is equal to the size of the input feature vector with n and k dependent on the size of m. To reduce model parameter tuning time, several feature vectors of different size m are used to test the optimal sizes of n and k for both the SVM and Random forest classifiers.

ElasticNet is a regularized regression method first proposed by Trevor Hastie[53]. ElasticNet combines the L1 penalties of Lasso regression and the L2 penalties of Ridge regression with the goal of performing as well as Lasso regression while overcoming the issues of Lasso when the number predictors is greater than the number of samples (p>n). Much like the Lasso method, ElasticNet simultaneously performs variable detection and continuous variable shrinkage. In addition, ElasticNet has also shown to group variables into subsets, which may be either included or excluded from the model depending on predictive capability. In addition to linear regression the ElasticNet method can also be used with logistic regression for classification. In this study logistic regression will be used with the malignant (1), and benign (0) classes. In addition, ElasticNet contains a tuning parameter used to specify the L1 to L2 ratio for predictor variable weighting. This study will evaluate a range of this ratio setting to determine the optimal predictive performance of the logistic regression model.

Classification

Three supervised learning classification methods with different learning approaches will be evaluated in this project: Support Vector Machine (SVM), Random Forests, and Logistic regression. Training and testing of the proposed classification methods will employ a k-fold cross-validation method. A more in-depth description of the k-fold cross-validation method used in this study is described in the *Model Training, Test, and Evaluation* section of this document.

Use of the SVM combined with texture features in both prostate cancer and brain tumor image analysis has been previously demonstrated[10,54,55]. The SVM is a supervised learning model that will be used to classify the two sets of mp-MRI data using texture feature statistics. The goal of the SVM is to construct a hyperplane that develops a maximum separation between the two classes, which in this case are malignant and benign. The SVM can support linear classification as well as non-linear classification using a "kernel" to map the inputs into higher-dimension feature space. Classification using the SVM has used several kernels: linear, polynomial, and Gaussian, in order to determine the best classification performance for the class distribution of this dataset. The kernel method takes data in the original space and projects it onto linear class separation boundaries in a higher dimension space. Generally, linear boundaries in the higher space have better training class separation and achieve better classification accuracy[56].

The Random Forest classification method combines an ensemble (forest) of de-correlated decision trees and uses a majority vote rule from each decision tree to determine class assignment. The bagging method (Boostrap aggregation) when used with random forests is used to grow the decision trees during training and take the average of the result. In the case of classification, bagging simply takes the majority vote of the trees in the forest. Essentially bagging reduces variance by averaging many noisy but unbiased models such as decision trees. Also, since the bagging method is used with a random forest ensemble of trees, we are less worried about overfitting any one tree since the overall classification is an average of all trees. Logistic regression uses a linear combination of input predictor variables to predict K classes, where K=2 in this study. The basis of logistic regression is to model the posterior probabilities of the given classes using linear functions of the input variables. In general, to find the best model, logistic regression attempts to fit by maximum likelihood (best fitting) using the conditional likelihood of the probability of a class K given input x. In this study, ElasticNet regression will be used as the penalty for the logistic regression classifier.

Model Training, Test, and Evaluation

As shown in Table 2, the image dataset available for training and testing of the classifiers in this study consists of 55 patients with malignant tumors providing 155 images, and 23 patients with benign masses providing 64 images. As can be seen, the dataset is unbalanced between the two classes, so as not to bias our predictions

toward malignant images, balancing the datasets as best as possible is of great importance. Adding both PZ and TZ images to the training and test datasets will expose the classifiers as evenly as possible to any differences that may exist between images from the different zones in another attempt to avoid any added bias.

Cross-validation will be used to estimate the prediction error of the various feature extraction, feature reduction, and classification method combinations. The cross-validation method breaks up the dataset into training and testing data, trains the classifier using the training data, and estimates the expected prediction error using the test data sample. Ideally the dataset would be large enough so that we could obtain the independent test data from the dataset and train with the remaining data. However, in this case, the dataset is too small to hold out large enough training and test sets from the original data. Therefore, the K-fold cross-validation method will be used. K-fold cross-validation divides the dataset into K folds of approximately equal size. During training and testing, one-fold is reserved for testing (calculation of prediction error), and the remaining K-1 folds are used for training (fitting the model). This training and test iteration are performed for k = 1, 2, …, K and the K estimates of prediction error are averaged.

A 10-fold cross validation is used in this study, so the malignant distribution will be 5-6 patients, and 15-16 images per fold. However, due to the unbalanced nature of the dataset, the benign distribution will be 2-3 patients, and 5-6 images per fold. In an attempt to best balance the dataset, the solution devised here is to duplicate the benign data between each pair of 5 folds as shown in Table 5, where M1 – M10 contain the 55 malignant patients and 155 associated images, and B1 – B5 contain the 23 benign patients and 64 associated images.

Fold 1	M1	B1
Fold 2	M2	B2
Fold 3	M3	B3
Fold 4	M4	B4
Fold 5	M5	B5
Fold 6	M6	B1
Fold 7	M7	B2
Fold 8	M8	B3
Fold 9	M9	B4
Fold 10	M10	B5

Table 5. 10-fold cross-validation splits

To avoid biasing the benign classes, the typical K-fold cross-validation method of holding one fold for testing and using the remaining folds for training, cannot be used with his method as a fold with the same benign data could be used for both testing and training. To avoid this bias, the fold that contains the same benign data in the training set, as that contained in the fold that is used for testing, will be held out and not used. In essence this makes the 10-fold cross-validation a 9-fold cross-validation. For example: if fold 1 is reserved for testing, fold 6 will be held out of the

training set thus making the training set fold (2,3,4,5,7,8,9,10). While not as desirable as having twice as much original benign data, this method will balance the dataset and avoid the unbalanced malignant bias seen in the 5-fold testing.

For each cross combination of extracted feature set, feature reduction method, and classifier, a Receiver Operating Characteristic (ROC) curve will be generated. The area under the ROC curve (AUC) will be used as the metric to judge a particular classifier's accuracy and used to compare classifiers and the extracted features used for that classifier. Note that when the term "classifier" is used in the context of accuracy or performance, the extracted feature and feature reduction method are implicitly included.

In addition to the AUC as a classifier performance measure, we consider the true positive (malignant) prediction accuracy, also known as sensitivity. In clinical terms, it is of the utmost importance not to send a patient with a malignant tumor home without a biopsy, even if this is at the expense of a biopsy performed on a benign patient. In other words, it is better to over biopsy than to under biopsy. This is not to say that we simply biopsy every patient with a suspicious looking MRI, but instead, develop a computer model that will attempt to identify all truly malignant patients while still identifying some, if not all, benign patients. Therefore, the sensitivity of our classifier should also be a top consideration when making comparisons of the various methods evaluated in this study.

Before evaluation of the 5 models used in this study, the various parameters of each model must be tuned. This tuning involves adjusting the parameters for a given dimensionality reduction/feature selection method and classifier combination to achieve the optimum classification accuracy. Extracted feature vectors are also required to tune the models presenting us with three variable sets: extracted features, dimensionality reduction/feature selection, and classification parameters. As one can see, the number of parameter values and the number of available feature vectors can grow quickly and make the optimization process computationally expensive and time consuming. To make the scope of the parameter optimization more manageable, subsets of each parameter value will be chosen to give a generally optimized model that can be used in the full model and feature evaluation process. For the classification software packages, the number of tuning parameters is pre-defined, for the dimensionality reduction/feature selection code the tuning parameters are method and architecture dependent. However, for the available extraction feature vectors, the number of vectors to use, and which extraction methods to use, is problem dependent. The strategy used here is to select a set of feature vectors that represent the size range of the number of extracted features. Table 6 lists the chosen extracted feature vectors and their sizes.

Extracted feature	Vector size
GLSZM, Bin size 16, (T2W)	13
MRFD, 4 directions (T2W and ADC)	32
GLCM, 4 directions, (ADC)	52
LBP, Bin size 32, (T2W and ADC)	72
GLRLM, 4 directions, Bin size 32, (T2W and ADC)	104
GLCM, LBP, MRFD, GLSZM, GLRLM (T2W)	169
GLCM, LBP, MRFD, GLSZM, GLRLM (T2W and ADC)	328

Table 6. List of seven representative tuning feature vectors.

The chosen seven representative extracted feature vectors are used to tune both the dimensionality reduction/feature selection methods and classifiers using the following tuning parameters:

mRMR

- Predictor correlation to class thresholds: [0.75, 0.5, 0.25].
- Intra-predictor correlation thresholds [0.15, 0.1, 0.05].
- All cross combinations of these thresholds will be tested.

Autoencoder. Rules for setting hidden and latent layers:

- Hidden layer size < Input and output layer size.
- Latent feature layer set to sizes: [5, 10, 20, 30]
- Hidden to latent layer ratios: [4:1, 3:1, 2:1]
- Epochs: [500, 1000, 1500, 2000, 2500, 3000]

SVM

- C: Penalty parameter of the error term. Set of [10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2} , 10^{3}]
- Polynomial degree (used only for polynomial kernel): [2, 3, 4, 5]
- Gamma (used for radial basis function kernel, it defines how far the influence of a single training example reaches): [1, 2, 3]
- Class_weight (Used for all kernels): "balanced" automatically determines class weights based on the class list given for training.

Random forest

- N_estimators (number of trees): [100, 250, 500, 1000]
- Criterion: gini
- Max_depth: None (nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples)
- Min_samples_leaf (The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least min_samples_leaf training samples in each of the left and right branches): [1, 5, 10, 15, 20]
- Min_samples_split (The minimum number of samples required to split an internal node): [10, 15, 20, 25, 30]

ElasticNet

- L1_ratio (ElasticNet mixing parameter, ratio of L1 to L2 penalty): [0.1, 0.2, 0.3, 0.4, 0.5]
- Max_iter: [100, 500, 1000]
- Loss: log
- Alpha: [0.1, 0.5, 1.0]
- Class_wieght: balanced

Software

All software run on MacBook Pro: 2.3 GHz Intel Core i7 CPU, 8 GB 1600 MHz DDR3 system memory, NVIDIA GeForce GT 650M, 512 MB graphics memory, macOS Sierra v10.12.6. Table 6 lists the software package and corresponding function used in this project.

Function	Software Package
ROI Extraction	Custom code written by author
Median kernel image de-noising	Custom code written by author
Variational Autoencoder image de-noising	Custom code written by author. Base code: https://github.com/keras-team/keras/blob/master/examples/variational_autoencoder.py
Convolutional Autoencoder image de-noising	Custom code written by author. Base code: John Ramey, https://ramhiser.com/post/2018-05-14-autoencoders-with-keras/
Feature Extraction (GLCM)	Mahotas version 1.4.4
Feature Extraction (LBP)	Mahotas version 1.4.4
Feature Extraction (Multi-resolution Fractal Dimensions)	Scikit-image version 0.14.0 and custom code written by author.
GLSZM	Custom code written by author
GLRLM	Custom code written by author
Dimensionality Reduction (ElasticNet)	Scikit-learn version 0.19.2
Dimensionality Reduction (Autoencoder)	Custom code written by author using the Keras front end version 2.1.6, and Tensorflow backend version 1.10.0
Feature Selection (mRMR)	Custom code written by author
Build k-folds for training and test	Custom code written by author
Classification (SVM)	Scikit-learn version 0.19.2
Classification (Random Forest)	Scikit-learn version 0.19.2
Classification (Logistic Regression)	Scikit-learn version 0.19.2
Ensemble Classification	Custom code written by author

Table 7. Software packages used and their function.

Chapter 3: Results

This section covers the results of the image de-noising methods, tuning of the feature dimensionality reduction/selection and classification methods, and the classification results of all models for all extracted features. In addition to presenting the classification results at the image level, the move to a patient level diagnosis decision support model is also discussed and results presented.

Image de-noising results

In the image pre-processing step, three image de-noising techniques were used on the original mp-MRI images. The question we have to answer is: which of the de-noising techniques, if any, result in the best prediction performance of our models? The first step in answering this question is to evaluate the quality of the resulting de-noised images by comparing them to the original image. If the de-noising technique degrades the pattern and information inherent in the original image to a degree in which any relevant feature information is lost, the images obviously cannot be used. Remember, the goal of image de-noising is to remove noise, which introduces bias and artifacts that have a negative impact on feature extractions and as a consequence on image classification accuracy. To evaluate the de-noised image deviation from the original image, mean squared error (MSE) was used. Table 8 shows the average of the MSE scores for all T2W and ADC images for the three de-noising techniques used. As can be seen, the VAE method vastly underperforms that of the Median kernel and CAE methods and was not used in assessing model performance using de-noised images. It is also shown that the Median kernel method outperforms the CAE method, however both methods will be used to assess model performance using de-noised images.

Image type	Kernel Method	VAE Method	CAE Method
T2W	12.9	366.3	27.5
ADC	8.6	1115.6	80.6

Table 8. Average MSE for T2W and ADC images for all de-noising techniques.

Figures 17 and 18 allow us to visually view the image differences between a representative original ROI image and the same image de-noised using the 3 methods.

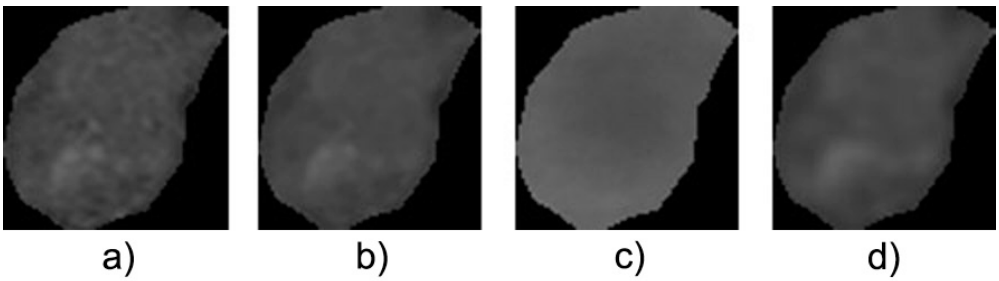

Figure 17. T2W original and de-noised ROI images. (a) Original ROI, (b) Median kernel de-noised ROI, (c) VAE de-noised ROI, (d) CAE de-noised ROI

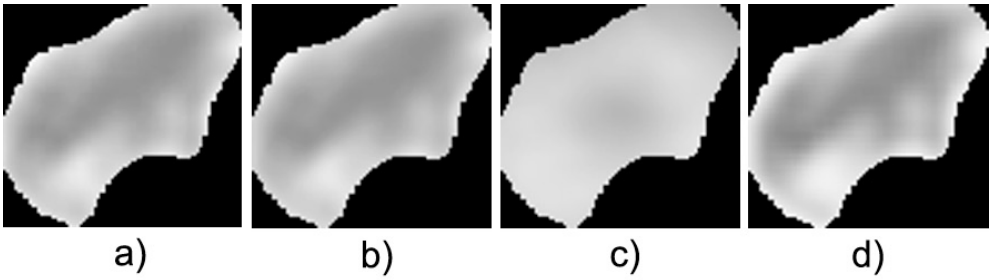

Figure 18. ADC original and de-noised ROI images. (a) Original ROI, (b) Median kernel de-noised ROI, (c) VAE de-noised ROI, (d) CAE de-noised ROI

As can be observed in Figure 17, the pattern of pixel intensities (light and dark areas), inherent in the original image (Figure 17a), are mostly preserved by the median filer (Figure 17b), and the CAE (Figure 17d). However, the VAE de-noised image has lost much of the pattern that existed in the original image, thus removing the significant texture patterns useful for classification. Figure 18 shows the same pattern corruption in the VAE de-noised ADC ROI image (Figure 18c) as compared to the original image (Figure 18). Table 8 shows the MSE comparison for the T2W and ADC representative ROI image shown in Figures 17 and 18. As can be seen Table 9, the VAE, as in Table 8, shows the same large amount of MSE error.

Image type	Kernel Method	VAE Method	CAE Method
T2W	3.9	268.6	7.1
ADC	3.6	1003.9	52.8

Table 9. MSE for representative T2W and ADC images for all de-noising techniques.

Having selected ROI images suitable for feature extraction, the second portion of the image de-noising question can be answered: which image ROI, original or de-noised, provides the highest classification accuracy? To answer this question, the original ROI images and each of the two selected de-noised ROI image types were

used in the analysis workflow using the models shown in Table 10. AUC scores were used as the metric for comparison showing the maximum (mean +/- SD) AUC for the top extracted feature for each model.

Analysis Model	Original ROI	Median kernel de-noised ROI	CAE de-noised ROI
mRMR + SVM	0.730 +/- 0.106	0.729 +/- 0.071	0.739 +/- 0.091
mRMR + Random Forest	0.754 +/- 0.154	0.739 +/- 0.103	0.725 +/- 0.101
Autoencoder + SVM	0.697 +/- 0.132	0.693 +/- 0.070	0.694 +/- 0.078
Autoencoder + Random Forest	0.685 +/- 0.097	0.655 +/- 0.073	0.694 +/- 0.072
ElasticNet	0.710 +/- 0.072	0.682 +/- 0.166	0.685 +/- 0.072

Table 10. Resulting AUC for each image type, de-noising method, and model. The maximum AUC is shown for the top scoring extracted feature. AUC is represented by mean +/- standard deviation.

Overall, there is not much difference shown in the maximum AUC between the three ROI image types. The CAE de-noised results do show a slight improvement over the median kernel method and the original image for 2 of the 5 models; however, this difference is not thought to be significant. The original images showed the best maximum AUC score for 3 of the 5 models and thus are the image types used in all further analysis. We can conclude that the noise present in the original images does not effect classification performance, likely because likely the feature extraction methods are robust to noise like LBP or GLCM/GLRLM/GLSZM after intensity binning, or for the de-noised images, noise was removed along with discriminative/crucial information that negatively impacts classifier performance. However, I recommend that if an image de-noising method is used in a study similar to this one, that the CAE method be implemented as it showed greater maximum AUC scores for the five models than did the kernel method, although the kernel method showed a better overall MSE than that of the CAE method.

Model tuning

Before any analysis of the models was performed, the parameters of the models were tuned for maximum classification accuracy using the test extracted feature set. The parameters chosen for the feature dimensionality reduction and feature selection methods, and all classifiers are:

mRMR

- Predictor correlation to class and inter-predictor correlation thresholds: [0.5, 0.05], [0.25, 0.1]

Autoencoder

- Hidden / Latent layer sizes: 60/20, 30/10

SVM

- C: 1.0
- Polynomial degree: 2
- Gamma (Radial basis function): 1
- Class_weight: balanced

Random forest

- N_estimators (number of trees): 500
- Criterion: gini
- Max_depth: None (nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples)
- Min_samples_leaf: 5
- Min_samples_split: 25

ElasticNet

- L1_ratio (Ratio of L1 to L2 penalty): [0.1, 0.2, 0.3, 0.4, 0.5]
- Max_iter: 1000
- Loss: log
- Alpha: 0.1
- Class_wieght: balanced

T2W and ADC Image Analysis

With the models tuned, features were extracted for both the T2W and ADC ROI images. At this step, the goal was to identify the top performing feature and model combinations for the T2W and ADC ROI images, using the classifier ROC curve AUC as the evaluation metric. The resulting top 10 features for each model are shown in Tables 11 – 15

Features from T2W	AUC (mean +/- SD)	Features from ADC	AUC (mean +/- SD)
GLRLM (Bin size 32, 4 directions)	0.688 +/- 0.093	GLCM (Bin size 1, average 4 directions), LBP (Bin size 1), GLRLM (Bin size 1, 4 directions)	0.695 +/- 0.140
GLRLM (Bin size 1, average 4 directions)	0.678 +/- 0.172	LBP (Bin size 1, normalized)	0.693 +/- 0.079
LBP (Bin size 16)	0.658 +/- 0.104	GLCM (Bin size 8, 4 directions)	0.682 +/- 0.163
GLRLM (Bin size 1, 4 directions)	0.646 +/- 0.162	GLCM (Bin size 1, average 4 directions)	0.677 +/- 0.166
LBP (Bin size 1, normalized)	0.644 +/- 0.084	GLCM (Bin size 16, 4 directions)	0.677 +/- 0.169
LBP (Bin size 8, normalized)	0.640 +/- 0.116	GLCM (Bin size 1, 4 directions)	0.674 +/- 0.164
LBP (Bin size 32, normalized)	0.640 +/- 0.105	MRFD (4 directions)	0.669 +/- 0.124
LBP (Bin size 1), GLRLM (Bin size 1, 4 directions)	0.616 +/- 0.188	MRFD (4 directions)	0.667 +/- 0.109
LBP (Bin size 32)	0.607 +/- 0.091	GLCM (Bin size 1, 4 directions), LBP (Bin size 1), GLRLM (Bin size 1, 4 directions)	0.665 +/- 0.138
GLSZM (Bin size 32)	0.605 +/- 0.089	MRFD (average 4 directions)	0.663 +/- 0.128

Table 11. T2W and ADC results for mRMR + SVM

In Table 11 it can be seen that for the mRMR + SVM model, neither T2W nor ADC outperforms the other for the top extracted feature, although the T2W AUC scores decrease faster than the ADC AUC scores for the top 10 features. It is also observed in Table 11 that the top 10 features from T2W consist mostly of LBP and GLRLM, whereas for ADC images, the top 10 AUC scores are dominated by the GLCM and MRFD.

T2W Feature type	AUC (mean +/- SD)	ADC Feature type	AUC (mean +/- SD)
LBP (Bin size 1), GLRLM (Bin size 1, 4 directions)	0.663 +/- 0.138	LBP (Bin size 8)	0.714 +/- 0.139
LBP (Bin size 1), GLRLM (Bin size 1, 4 directions), MRFD (4 directions)	0.663 +/- 0.137	LBP (Bin size 16)	0.680 +/- 0.113
GLRLM (Bin size 1, 4 directions)	0.651 +/- 0.138	LBP (Bin size 1)	0.659 +/- 0.111
GLRLM (Bin size 1, average 4 directions)	0.629 +/- 0.126	LBP (Bin size 1), GLRLM (Bin size 1, 4 directions), MRFD (4 directions)	0.659 +/- 0.110
LBP (Bin size 1, normalized)	0.617 +/- 0.084	MRFD (4 directions)	0.658 +/- 0.089
LBP (Bin size 32)	0.609 +/- 0.083	LBP (Bin size 32)	0.656 +/- 0.101
GLCM (Bin size 1, average 4 directions), LBP (Bin size 1), GLRLM (Bin size 1, 4 directions)	0.604 +/- 0.138	GLSZM (Bin size 32)	0.655 +/- 0.089
LBP (Bin size 8)	0.594 +/- 0.108	LBP (Bin size 1), GLRLM (Bin size 1, 4 directions)	0.653 +/- 0.087
GLCM (Bin size 1, 4 directions), LBP (Bin size 1), GLRLM (Bin size 1, 4 directions)	0.594 +/- 0.150	GLRLM (Bin size 16, 4 directions)	0.618 +/- 0.131
LBP (Bin size 32, normalized)	0.585 +/- 0.106	GLCM (Bin size 8, 4 directions)	0.613 +/- 0.140

Table 12. T2W and ADC results for mRMR + Random Forest

Comparing the best results for the mRMR + Random forest model in the top row of Table 12, it is observed that the ADC feature significantly outperforms the T2W feature with an AUC of 0.754 vs. 0.663. The trend of ADC outperforming T2W continues for the top 10 extracted features. For this model, the top 10 features for both T2W and ADC show more diversity than seen for the mRMR + SVM model.

T2W Feature type	AUC (mean +/- SD)	ADC Feature type	AUC (mean +/- SD)
LBP (Bin size 16)	0.664 +/- 0.071	MRFD (4 directions)	0.661 +/- 0.116
LBP (Bin size 1)	0.650 +/- 0.144	GLCM (Bin size 16, 4 directions)	0.658 +/- 0.157
LBP (Bin size 16, GLRLM (Bin size 1, 4 directions)	0.641 +/- 0.090	GLCM (Bin size 8, 4 directions)	0.650 +/- 0.146
LBP (Bin size 32, Normalized)	0.619 +/- 0.117	MRFD (average 4 directions)	0.636 +/- 0.111
GLRLM (Bin size 1, average 4 directions)	0.606 +/- 0.143	GLCM (Bin size 1, 4 directions)	0.622 +/- 0.154
GLRLM (Bin size 1, 4 directions)	0.601 +/- 0.097	GLRLM (Bin size 1, 4 directions)	0.620 +/- 0.108
LBP (Bin size 8)	0.595 +/- 0.064	GLSZM (Bin size 8)	0.620 +/- 0.189
LBP (Bin size 1, normalized)	0. 582 +/- 0.161	GLCM (Bin size 1, average 4 directions	0.616 +/- 0.169
LBP (Bin size 32)	0.570 +/- 0.138	GLRLM (Bin size 1, average 4 directions	0.615 +/- 0.101
LBP (Bin size 16), GLRLM (Bin size 16, 4 directions), MRFD (4 directions)	0.568 +/- 0.115	GLSZM (Bin size 32)	0.606 +/- 0.078

Table 13. T2W and ADC results for Autoencoder + SVM

Like the mRMR + SVM model shown in Table 11, the autoencoder + SVM model shows no performance difference between T2W and ADC features. For this model however, the T2W top scoring features are dominated almost exclusively by the LBP extraction method while the ADC image features show a variance in feature extraction type.

T2W Feature type	AUC (mean +/- SD)	ADC Feature type	AUC (mean +/- SD)
LBP (Bin size 16, GLRLM (Bin size 1, 4 directions)	0.640 +/- 0.104	MRFD (4 directions)	0.645 +/- 0.150
LBP (Bin size 16, GLRLM (Bin size 16, 4 directions)	0.636 +/- 0.098	GLRLM (Bin size 1, 4 directions)	0.613 +/- 0.116
LBP (Bin size 1)	0.625 +/- 0.143	MRFD (average 4 directions)	0.607 +/- 0.119
LBP (Bin size 32)	0.621 +/- 0.078	LBP (Bin size 32, normalized)	0.605 +/- 0.094
LBP (Bin size 16)	0.599 +/- 0.085	GLRLM (Bin size 1, average 4 directions)	0.603 +/- 0.109
LBP (Bin size 8, normalized)	0.594 +/- 0.083	LBP (Bin size 32)	0.599 +/- 0.060
GLRLM (Bin size 1, average 4 directions)	0.557 +/- 0.184	MRFD (4 directions)	0.596 +/- 0.107
LBP (Bin size 8)	0.556 +/- 0.057	GLCM (Bin size 16, 4 directions)	0.596 +/- 0.149
GLCM (Bin size 1, 4 directions)	0.547 +/- 0.096	GLSZM (Bin size 32)	0.595 +/- 0.088
GLRLM (Bin size 8, 4 directions)	0.536 +/- 0.153	LBP (Bin size 16)	0.589 +/- 0.076

Table 14. T2W and ADC results for Autoencoder + Random Forest

The results for the autoencoder + random forest model shown in Table 14 parallel those shown for the autoencoder + SVM in Table 13: no performance difference between T2W and ADC features, the T2W top scoring features are dominated almost exclusively by the LBP extraction method while the ADC image features show a variance in feature extraction type.

T2W Feature type	AUC (mean +/- SD)	ADC Feature type	AUC (mean +/- SD)
LBP (Bin size 32)	0.646 +/- 0.088	GLCM (Bin size 8, 4 directions)	0.679 +/- 0.152
LBP (Bin size 16)	0.636 +/- 0.095	GLCM (Bin size 16, 4 directions)	0.678 +/- 0.151
LBP (Bin size 16) GLCM (Bin size 16, 4 directions)	0.634 +/- 0.072	GLCM (Bin size 1, 4 directions)	0. 668 +/- 0.164
GLSZM (Bin size 32)	0.614 +/- 0.118	MRFD (0 deg)	0.659 +/- 0.117
LBP (Bin size 32, normalized)	0.598 +/- 0.086	GLCM (Bin size 1, average 4 directions)	0.657 +/- 0.161
GLRLM (Bin size 32, 4 directions)	0.597 +/- 0.146	MRFD (average 0, 45 90, 135 deg)	0.656 +/- 0.110
LBP (Bin size 8, normalized)	0.586 +/- 0.149	MRFD (0, 45 90, 135 deg)	0.645 +/- 0.124
LBP (Bin size 1, normalized)	0.575 +/- 0.110	GLRLM (Bin size 1, 4 directions	0.626 +/- 0.151
LBP (Bin size 8)	0.550 +/- 0.157	LBP (Bin size 32)	0.614 +/- 0.087
LBP (Bin size 16, normalized)	0.548 +/- 0.139	LBP (Bin size 16)	0.613 +/- 0.083

Table 15. T2W and ADC results for ElasicNet

Once again for the results of the ElasticNet model shown in Table 15, no performance difference between T2W and ADC features is shown and the T2W top scoring features are dominated almost exclusively by the LBP extraction method while the ADC image features show a variance in feature extraction type.

Several conclusions can be made comparing the performance of the T2W and ADC image classification scores over all five model types and extracted features. First, neither T2W nor ADC significantly outperforms the other for the top AUC score in 3 of the 5 models, and ADC slightly outperforms T2W in two of the models. Thus, it can be concluded that neither modality has a distinct classification accuracy advantage in over the other. Secondly, in all models, the ADC scores are reduced at a significantly less rate as compared to the T2W scores. It is because of the more pronounced reduction in AUC score for the T2W images that only the top 5 scoring image features for T2W and ADC will be used to make multi-modal feature combinations.

Another interesting trend is seen in the feature types corresponding to the T2W and ADC modalities. The LBP feature extraction method is strongly associated with T2W images across all models, where a more diverse set of extracted features is associated with the ADC images. In fact, LBP will also play a role in the highest classification AUC scores when multi-modal features are analyzed.

Complete feature and model analysis

With the model analysis complete for each single image modality, multi-parametric analysis of each model was performed. Two sets of multi-modal features (features from both T2W and ADC images) were used in model evaluation: those obtained through the cross-combination of the top 5 AUC scoring features from each model, and the combination of T2W and ADC features from the same feature extraction method. The two combinations of T2W and ADC modalities are written as: T2W + ADC, when the same feature extraction method is used for both modalities, and T2W / ADC, when different feature extraction methods are used for each modality. The results of the top 10 performing extracted features for each of the five models are shown in Tables 16 – 20 for each of the five models used for classification. In each table, the AUC score selects the top 10 features for that particular model. Also in each table, several other classification metrics are listed:

- Sensitivity or true positive rate, the proportion of true positive predictions, which, in this study is the malignant class.
- Specificity or true negative rate, the proportion of true negatives that are classified as negative, which in this study is the benign class.
- Positive prediction value (PPV), which is the portion of true positives vs. false positives.
- Negative prediction rate (NPV), which gives the proportion of true negative predictions vs. false negative predictions.

Feature	Modality	AUC Mean +/- SD	Sensitivity	Specificity	PPV	NPV
LBP (Bin size 1)	T2W + ADC	0.730 +/- 0.106	0.88	0.44	0.66	0.76
GLCM(Bin size 1, Ave 4 directions), LBP (Bin size 1), GLRLM (Bin size 1, 4 directions)	T2W + ADC	0.728 +/- 0.114	0.82	0.55	0.69	0.71
T2W+ADC: LBP (Bin size 16), ADC: GLRLM (Bin size 16, 4 directions)	T2W / ADC	0.724 +/- 0.094	0.74	0.66	0.72	0.67
LBP (Bin size 1), GLRLM (Bin size 1, 4 directions)	T2W + ADC	0.710 +/- 0.110	0.90	0.42	0.65	0.78
T2W: LBP(Bin size 32, Normalized), ADC: LBP(Bin size 16, Normalized)	T2W / ADC	0.710 +/- 0.114	0.78	0.48	0.65	0.65
LBP (Bin size 16)	T2W + ADC	0.708 +/- 0.094	0.69	0.59	0.67	0.61
GLCM(Bin size 1, 4 directions), LBP (Bin size 1), GLRLM (Bin size 1, 4 directions)	T2W + ADC	0.708 +/- 0.101	0.81	0.56	0.69	0.71
GLCM(Bin size 1, Ave 4 directions), LBP (Bin size 1), GLRLM (Bin size 1, 4 directions)	ADC	0.695 +/- 0.141	0.72	0.55	0.66	0.62
LBP (Bin size 1, Normalized)	ADC	0.693 +/- 0.079	0.85	0.37	0.62	0.67
GLRLM (Bin size 32, 4 directions)	T2W	0.688 +/- 0.093	0.74	0.52	0.65	0.62

Table 16. Multiparametric results for mRMR + SVM

Feature	Modality	AUC Mean +/- SD	Sensitivity	Specificity	PPV	NPV
T2W: LBP (Bin size 1), ADC: LBP (Bin size 8)	T2W / ADC	0.754 +/- 0.154	0.80	0.58	0.70	0.70
LBP (Bin size 8)	T2W + ADC	0.737 +/- 0.102	0.83	0.50	0.67	0.70
LBP (Bin size 1)	T2W + ADC	0.721 +/- 0.096	0.82	0.52	0.67	0.70
T2W: LBP (Bin size 1, normalized), ADC: LBP (Bin size 8)	T2W / ADC	0.721 +/- 0.098	0.83	0.52	0.67	0.71
T2W+ADC: LBP (Bin size 8), T2W: GLRLM (Bin size 1, 4 deg)	T2W + ADC	0.720 +/- 0.111	0.77	0.55	0.67	0.66
LBP (Bin size 8)	ADC	0.714 +/- 0.139	0.83	0.48	0.66	0.70
T2W: GLRLM (bin size 1, 4 directions), ADC: LBP (Bin size 8)	T2W / ADC	0.713 +/- 0.129	0.78	0.54	0.67	0.67
ADC: LBP (Bin size 8), ADC: MRFD (4 directions)	ADC / ADC	0.707 +/- 0.146	0.77	0.55	0.67	0.66
LBP (Bin size 16)	T2W + ADC	0.701 +/- 0.117	0.75	0.51	0.65	0.63
LBP (Bin size 1), GLRLM (Bin size 1, 4 directions), MRFD (4 directions)	T2W + ADC	0.697 +/- 0.114	0.79	0.50	0.66	0.66

Table 17. Multiparametric results for mRMR + Random Forest

Feature	Modality	AUC Mean +/- SD	Sensitivity	Specificity	PPV	NPV
LBP (Bin size 16, Normalized)	T2W + ADC	0.697 +/- 0.132	0.83	0.39	0.62	0.65
LBP (Bin size 16)	T2W + ADC	0.677 +/- 0.133	0.88	0.22	0.58	0.61
LBP (Bin size 16)	T2W	0.661 +/- 0.116	0.75	0.38	0.59	0.55
LBP (Bin size 16), MRFD (4 directions)	ADC	0.658 +/- 0.157	0.75	0.47	0.63	0.61
LBP (Bin size 16, Normalized)	T2W + ADC	0.655 +/- 0.124	0.88	0.22	0.58	0.61
GLCM (Bin size 16, 4 directions)	ADC	0.650 +/- 0.124	0.92	0.11	0.55	0.52
MRFD (4 directions)	T2W + ADC	0.650 +/- 0.146	0.74	0.47	0.63	0.60
GLCM (Bin size 8, 4 directions)	ADC	0.650 +/- 0.144	0.76	0.41	0.61	0.59
LBP (Bin size 1)	T2W	0.650 +/- 0.127	0.74	0.39	0.60	0.56
MRFD (4 directions)	T2W + ADC	0.650 +/- 0.124	0.75	0.38	0.59	0.55

Table 18. Multiparametric results for Autoencoder + SVM

Feature	Modality	AUC Mean +/- SD	Sensitivity	Specificity	PPV	NPV
LBP (Bin size 16), GLRLM (Bin size 16), MRFD (4 directions)	T2W + ADC	0.685 +/- 0.097	0.77	0.47	0.64	0.63
T2W: LBP (Bin size 16), ADC: GLRLM (Bin size 16, 4 directions)	T2W / ADC	0.646 +/- 0.084	0.73	0.38	0.59	0.54
MRFD (4 directions)	ADC	0.645 +/- 0.150	0.75	0.45	0.62	0.60
LBP (Bin size 16), GLRLM (Bin size 1, 4 directions)	T2W	0.640 +/- 0.104	0.72	0.30	0.56	0.48
LBP (Bin size 16), GLRLM (Bin size 16, 4 directions)	T2W	0.636 +/- 0.098	0.71	0.30	0.55	0.46
LBP (Bin size 16, Normalized)	T2W + ADC	0.632 +/- 0.131	0.72	0.40	0.59	0.54
LBP (Bin size 1)	T2W	0.625 +/- 0.114	0.76	0.44	0.62	0.60
LBP (Bin size 16), GLRLM (Bin size 16, 4 directions)	T2W + ADC	0.622 +/- 0.124	0.78	0.55	0.68	0.68
LBP (Bin size 16), GLRLM (Bin size 8, 4 directions)	T2W + ADC	0.621 +/- 0.121	0.80	0.52	0.67	0.68
LBP (Bin size 32, Normalized)	T2W + ADC	0.621 +/- 0.099	0.75	0.30	0.55	0.51

Table 19. Multiparametric results for Autoencoder + Random forest

Feature	Modality	AUC Mean +/- SD	Sensitivity	Specificity	PPV	NPV
LBP (Bin size 16)	T2W + ADC	0.710 +/- 0.072	0.58	0.70	0.70	0.58
LBP (Bin size 16), GLCM (Bin size 16, 4 directions)	T2W + ADC	0.701 +/- 0.124	0.67	0.55	064	0.58
LBP (Bin size 8)	T2W + ADC	0.692 +/- 0.115	0.52	0.67	0.66	0.63
T2W+ADC: LBP (Bin size 16), ADC: GLCM (Bin size 8, 4 directions)	T2W / ADC	0.686 +/- 0.090	0.61	0.61	0.65	0.56
T2W: LBP (Bin size 16), ADC: GLCM (Bin size 8)	T2W / ADC	0.683 +/- 0.140	0.67	0.44	0.59	0.52
GLCM (Bin size 8, 4 directions)	ADC	0.679 +/- 0.152	0.40	0.77	0.68	0.51
GLCM (Bin size 16, 4 directions)	T2W + ADC	0.678 +/- 0.152	0.47	0.70	0.64	0.51
GLCM (Bin size 8, 4 directions)	T2W + ADC	0.678 +/- 0.152	0.52	0.64	0.64	0.53
GLCM (Bin size 16, 4 directions)	ADC	0.677 +/- 0.115	0.45	0.80	0.69	0.53
T2W: LBP (Bin size 16), ADC: GLRLM (Bin size 1, 4 directions)	T2W / ADC	0.675 +/- 0.082	0.65	0.63	0.68	0.59

Table 20. Multiparametric results for ElasticNet

From Tables 16-20 several observations and conclusions can be made. First, the best image level model is mRMR + Random forest model with the highest AUC classification accuracy of 0.754 +/- 0.154 when using the multiparametric extracted feature consisting of T2W: LBP (Bin size 1) and ADC: LBP (Bin size 8). A second observation can be made that LBP is totally, or partially, involved in the 5 models top AUC scores. A third observation can be made from the top AUC scores for the SVM and Random Forest classifiers using the mRMR and autoencoder dimensionality reduction methods; given the same classifiers, mRMR outperforms the autoencoder method.

Having determined the mRMR + Random Forest model is my recommendation for best image level model, the ROC curves for this model are shown in Figure 19 for all 10 folds used during training and testing of the model. The highlighted blue curve in Figure 19 represents the mean of the 10-fold curves, and the grey area representing the standard deviation.

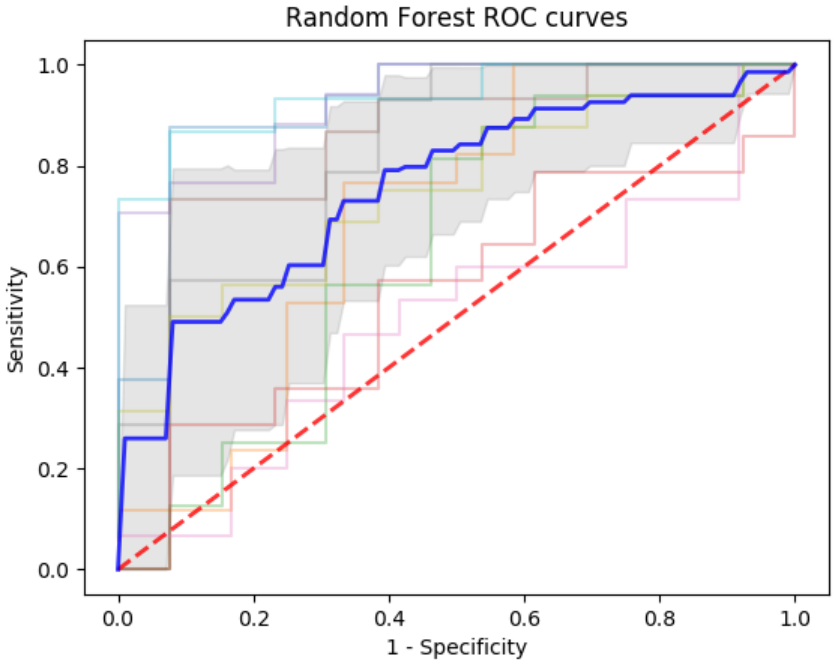

Figure 19. ROC curves for the mRMR + Random forest model using combined T2W and ADC LBP features. Mean AUC (highlighted blue line) of 0.754 +/- 0.154

A very important observation can be made from the results shown in Tables 16-20; multi-modal features account for at least the top two AUC scores for each of the 5 models. We can draw the conclusion that a multi-modal image analysis approach produces higher classification accuracy (AUC) than simply T2W or ADC images alone, and based on this conclusion, I recommend multiparametric features be used in MRI prostate cancer image analysis. To highlight this conclusion, the features for the top 2 AUC scores for the 5 models are shown in Table 21. The T2W and ADC columns show the individual image modality AUC while the multiparametric column shows that for all models, the sum of the individual image types is greater than either type alone.

Model	Feature	T2W (mean +/- SD)	ADC (mean +/- SD)	Multiparametric (mean +/- SD)
mRMR + SVM	LBP (Bin size 1)	0.552 +/- 0.118	0.657 +/- 0.119	0.730 +/- 0.016
mRMR + SVM	GLCM(Bin size 1, Ave 4 directions), LBP (Bin size 1), GLRLM (Bin size 1, 4 directions)	0.572 +/- 0.172	0.634 +/- 0.162	0.728 +/- 0.114
mRMR + Random forest	T2W: LBP (Bin size 1), ADC: LBP (Bin size 8)	0.568 +/- 0.091	0.659 +/- 0.110	0.754 +/- 0.154
mRMR + Random forest	LBP (Bin size 8)	0.594 +/- 0.108	0.714 +/- 0.139	0.737 +/- 0.102
Autoencoder + SVM	LBP (Bin size 16, Normalized)	0.487 +/- 0.141	0.586 +/- 0.110	0.697 +/- 0.132
Autoencoder + SVM	LBP (Bin size 16)	0.567 +/- 0.103	0.475 +/- 0.087	0.677 +/- 0.133
Autoencoder + Random forest	LBP (Bin size 16), GLRLM (Bin size 16), MRFD (4 directions)	0.541 +/- 0.106	0.529 +/- 0.147	0.685 +/- 0.097
Autoencoder + Random forest	T2W: LBP (Bin size 16), ADC: GLRLM (Bin size 16, 4 directions)	0.599 +/- 0.085	0.573 +/- 0.097	0.646 +/- 0.084
ElasticNet	LBP (Bin size 16)	0.630 +/- 0.080	0.613 +/- 0.083	0.710 +/- 0.072
ElasticNet	LBP (Bin size 16), GLCM (Bin size 16, 4 directions)	0.634 +/- 0.073	0.597 +/- 0.113	0.701 +/- 0.124

Table 21. Comparison of single and multiparametric image modalities. The top 2 AUC scores for each of the 5 models are shown.

Using a multiparametric feature analysis approach generates larger extracted feature predictor vectors than a single mode analysis approach, and puts more emphasis on dimensionality reduction and feature selection methods to find the best features for optimum class prediction. The mRMR feature selection method allows us to not only reduce the size of the feature predictor vector, but also directly identify the features selected. For the mRMR + SVM models top AUC feature set, 72 features are generated by the LBP feature extraction method, of which, 15 features

are selected as the input to the classifier. Of these 15 features, 5 are selected from the T2W ROI image and 10 from the ADC ROI mage. Similarly, for the mRMR + Random forest models top AUC feature set, 72 features are generated by the LBP feature extraction method of which 9 features are selected for the reduced inputs to the classifier. Of these 9 features, 1 is selected from the T2W ROI image and 8 are from the ADC ROI mage. The two examples here show that the ADC extracted features for the mRMR method correlate better to the classifier than do the T2W extracted features. The autoencoder also generates a reduced size encoded latent feature vector from the original input, however unlike the mRMR method the reduced feature vector cannot easily be matched to the original features. For the autoencoder + SVM model, 72 feature predictors and reduced to an encoded latent variable of 20 features. The autoencoder + Random forest model reduces an input extracted feature vector of 208 features to an encoded latent feature vector of size 20.

From Tables 16 – 20, we can get an idea of the classification accuracy from the AUC, sensitivity, specificity, PPV, and NPV of the five models evaluated. However, it is also helpful in understanding the various strengths of each model. As can be seen in Tables 16 – 20, classification accuracy metrics vary between models. Figure 20 shows a graphical representation of the 5 classification metrics for the 5 classification models top extracted features.

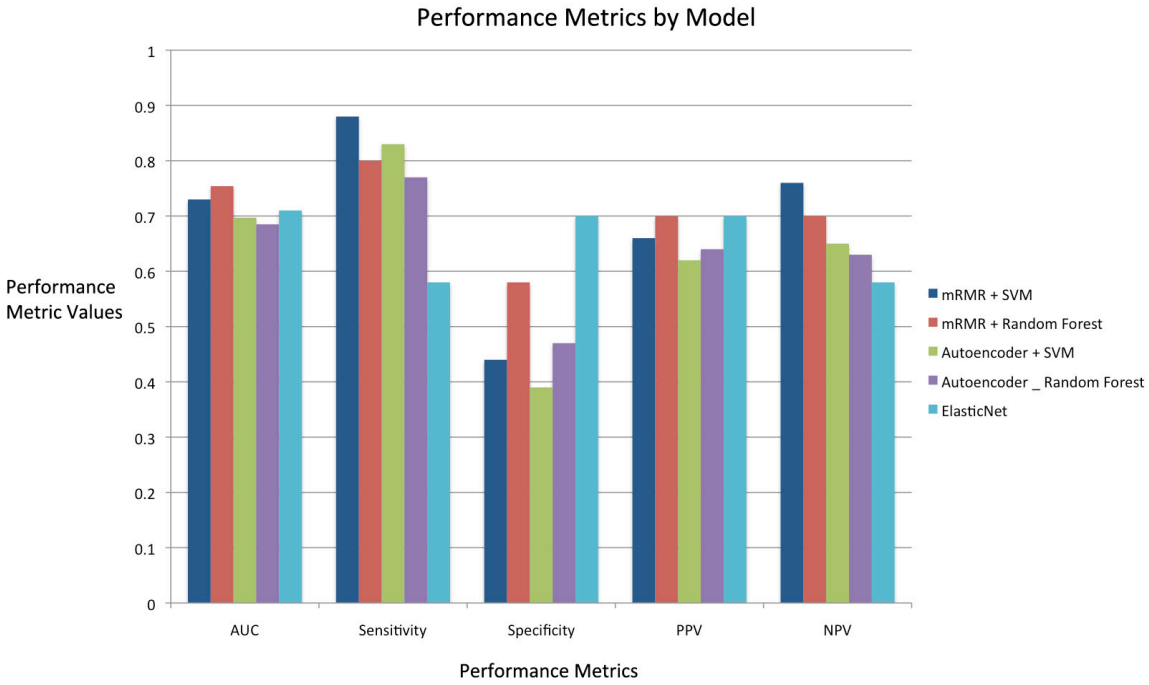

Figure 20. Performance metrics for all 5 classification models.

As can be seen, although the mRMR + Random Forest model has the highest AUC score (0.754), it does not have the highest sensitivity, and shows a tradeoff between sensitivity and specificity. Comparing the mRMR + Random forest model to the mRMR + SVM model which has the second highest AUC of 0.730, the mRMR + SVM

model a higher, and in fact the best sensitivity score. Once again the mRMR + SVM model trades a higher sensitivity for lower specificity, which is the opposite of that shown by the ElasticNet model. From Figure 19 it is shown that each model has different strengths and weaknesses that lead to the conclusion that an ensemble of classifiers can improve classification accuracy over any individual model.

Ensemble of Models Classification

To reduce the classification variance observed between the five individual models, an ensemble classification model was created using all five individual models previously discussed, and majority vote used to determine the final image classification. The confusion matrix for the ensemble classification model is shown in Table 22.

	Predicted Benign	Predicted Malignant
Expected Benign	66	62
Expected Malignant	25	130

Table 22. Image level ensemble model confusion matrix.

The AUC for the ensemble model is 0.773 +/- 0.070 and the ROC curves shown in Figure 21. A sensitivity of 0.84, and specificity of 0.52 for the ensemble model are comparable to the mRMR + Random forest model (0.80 and 0.58).

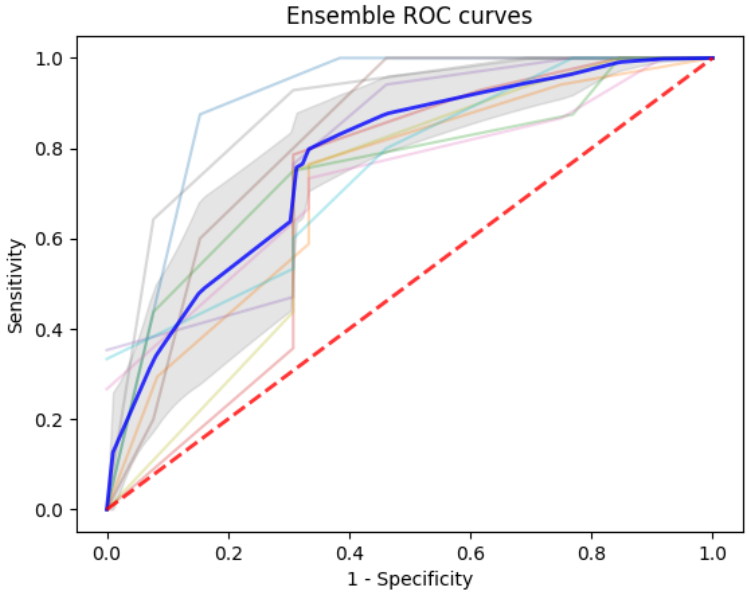

Figure 21. ROC curves for ensemble classifier. Mean AUC of 0.773 shown in blue, standard deviation shown in gray.

In order to observe how the ensemble model compares over the 5 classification metrics, these metrics are shown for each model in Figure 22.

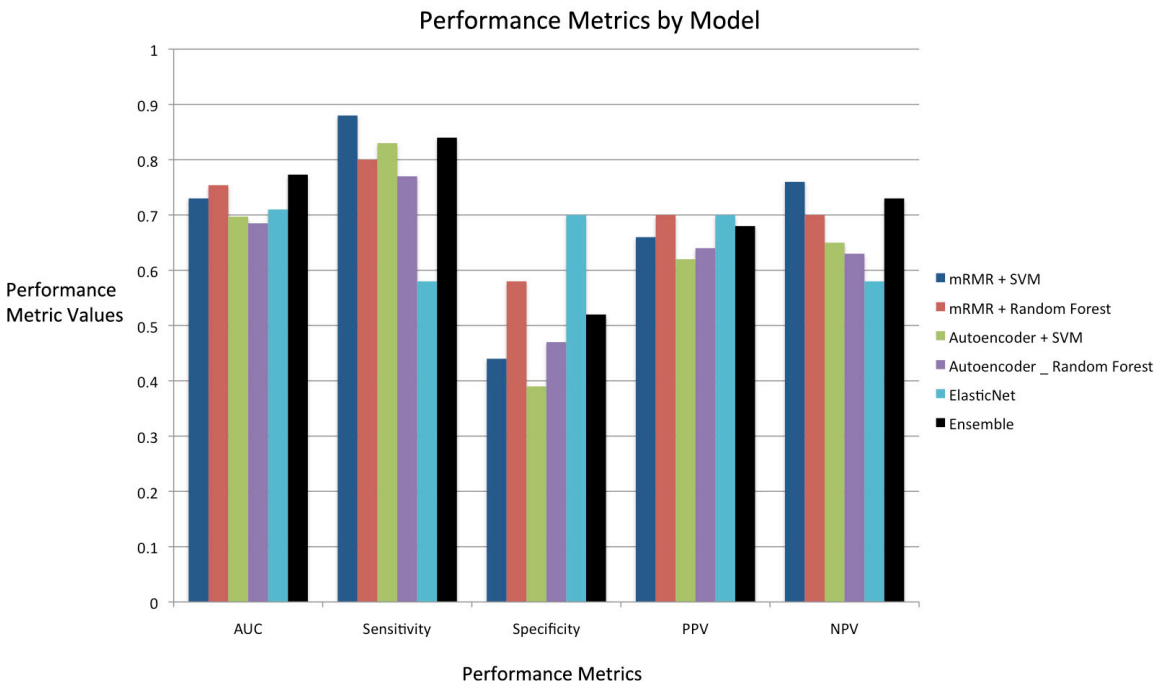

Figure 22. Performance metrics for all 6 classification models

Figure 22 shows that besides the ensemble model (shown with the black bar) having the largest AUC, it scores at or near the top for 4 of the 5 classifiers, with specificity being the lower scoring metric. This graph demonstrates the ensemble of classifiers does take advantage of the strengths of the individual models to build the best image level classification model for this study.

From the confusion matrix in Table 22, it can be seen that benign image slices are misclassified at almost a 2.5:1 compared to the malignant image slices. This is also observed in the other models with the exception of ElasticNet. From the ensemble model however, more information on which images are misclassified for each model relative to the others can be observed. In taking the majority vote for the ensemble classifier, one benign labeled patient stood out from the others. For this patient, all 3 images in the set are misclassified as malignant over all models, leading one to suspect that the patient class may have been mislabeled when the original dataset was compiled. A comparison is made in Figure 23 between the T2W and ADC ROI images for the suspected mislabeled patient, a second benign patient with 3 images, and a malignant patient with 3 images, all from the same fold used for classification.

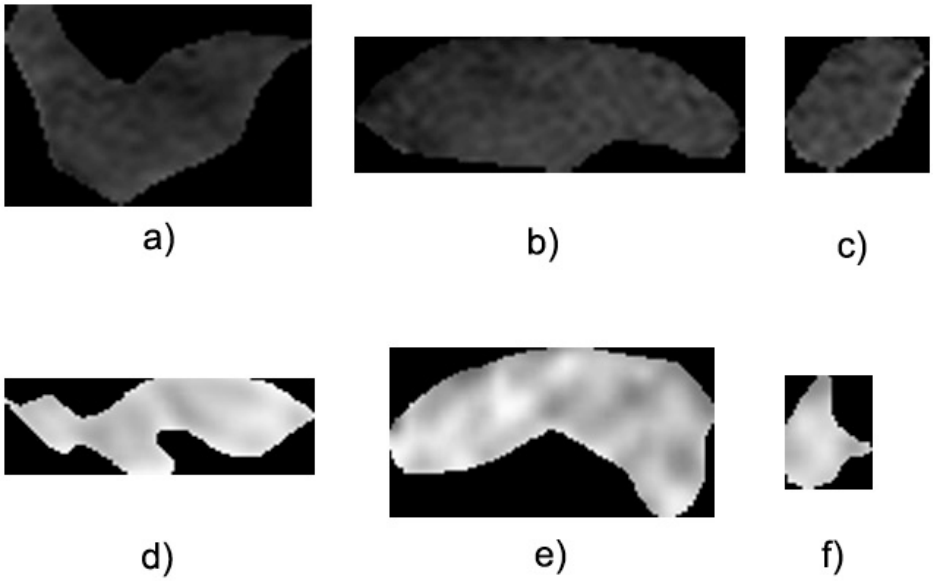

Figure 23. T2W and ADC ROI images for comparing misclassified benign patient. Misclassified benign patient (a, d), malignant classified patient (b, e), and benign classified patient (c, f).

Although human viewing and classification of images is always subjective, in my opinion the T2W images in Figure 23 a), b), and c) look very similar with no correlation between the potentially mislabeled image and the properly classified benign and malignant images. The ADC ROI image in Figure 23e) does look somewhat different than those in d) and f), possibly indicating that the patient is properly labeled as benign. I believe however, that based on the visual analysis of the images in Figure 23, no conclusive evidence exists to support calling the patient in question mislabeled.

To this point, a great deal of results information has been presented and I believe a brief overview of the results is warranted. It has been shown that the original images with no de-noising provided the best model classification accuracy. However, the CAE method performed well, and if de-noising is thought to be need on an image dataset, the CAE method is recommended. In the next step, T2W and ADC image extracted features were assessed for model classification accuracy and the results showed no significant difference between the image modalities. However, multiparametric features showed much better classification accuracy than either single image feature supporting the hypothesis that mp-MRI is the preferred extracted feature for a prostate cancer image analysis study as defined in this work. Using multiparametric extracted image features, it was shown that the mRMR + Random Forest was the best single classifier model, but that the best overall performing image level model was the ensemble of classifiers.

Patient level classification

Now that the image classification performance of all models has been evaluated, the next question is: how can these models be used in a clinical setting? To answer this question, we must move from image level classification, to patient level diagnosis, in order to produce a viable clinical diagnosis decision support model to aid in patient diagnosis. The goal of the patient level model is defined as: the patient level diagnosis model zero, or the lowest possible number of misdiagnosed malignant. The rational behind this goal is that it is better to over biopsy a benign patient than to misdiagnose a malignant patient with an aggressive cancer that goes untreated.

In order to determine the best patient level clinical model, a criterion needs to be established to differentiate between malignant and benign patients from the classification of their individual MRI images. In this analysis, a patient is considered malignant if one or more images for that patient have been classified as malignant. The reasoning behind only using one or more images to classify a patient as malignant is; because all images in a patient's image set are labeled based on biopsy results and are not correlated to a specific biopsy area, the image set may contain both benign and malignant images possibly based on intra-tumor heterogeneity. We should remember that in this study, the images in a patient's dataset are used to identify a suspected cancerous region and to define an ROI to guide biopsy and are not associated with a specific biopsy site. An assumption is made here that when using this model, that new patient images are acquired using the same method as those in the dataset used to train the classifiers; that is, regions of an image have been identified by a radiologist as possibly cancerous and these regions (ROI) outlined.

Using the criteria set above for patient level diagnosis, Table 23 shows the 6 classification models using their top scoring extracted features, using a malignant probability threshold of 0.5 to define a patient's diagnosis. Both properly diagnosed and misdiagnosed patients are shown. As a note, a misdiagnosed malignant patient is one that has it truly malignant and has been diagnosed as benign. This hold true for benign patient diagnosis as well.

Model	Misdiagnosed Malignant	Diagnosed Malignant	Misdiagnosed Benign	Diagnosed Benign
mRMR + SVM	1	54	14	9
mRMR + Random Forest	7	48	14	9
Autoencoder + SVM	3	52	19	4
Autoencoder + Random Forest	5	50	16	7
ElasticNet	17	38	12	11
Ensemble	3	52	15	8

Table 23. Patient diagnosis using malignant probability threshold of 0.5

Although the ensemble model has the highest AUC of all the models, it does not have the lowest number of misclassified malignant patients. In fact, the mRMR + SVM model has the lowest number of misclassified patients while having a reasonable benign patient misclassification as compared with the other models. To meet the goal of the lowest malignant misclassification, the mRMR + SVM model is the proper choice for the clinical diagnosis decision support model when using a malignant probability threshold of 0.5. However, a malignant threshold other than 0.5 can be chosen to evaluate patient diagnosis. Since our paramount goal is zero misdiagnosed malignant patients, even at the expense of misdiagnosed and over-biopsied truly benign patients, some error, or over confidence can be added by shifting the malignant probability threshold lower. In other words, we can state that we want to be absolutely sure a patient is benign. If we shift the malignant probability to 0.4, we say benign patients must have a probability of 0.6 before we diagnose them as such. Table 24 shows the patient diagnosis results when we shift the malignant probability threshold to 0.4.

Model	Misdiagnosed Malignant	Diagnosed Malignant	Misdiagnosed Benign	Diagnosed Benign
mRMR + SVM	0	55	15	8
mRMR + Random Forest	5	50	14	9
Autoencoder + SVM	0	55	19	4
Autoencoder + Random Forest	5	50	16	7
ElasticNet	3	52	12	11
Ensemble	1	54	15	8

Table 24. Patient diagnosis using malignant probability threshold of 0.4

From Table 24 we now can see two models (mRMR + SVM and autoencoder + SVM) that meet our goal of zero misdiagnosed malignant patients, but which model is better? In looking at the diagnosed benign patients between the two models, it can be seen that the mRMR + SVM model has 8 properly diagnosed benign patients vs. only 4 for the autoencoder + SVM model. The mRMR + SVM achieves the goal of zero misdiagnosed malignant patients and the best diagnosis of benign patients and is thus my recommended clinical diagnosis decision support model.

Chapter 4: Discussion

In this study I have elaborated a valid algorithm to use extracted features from mp-MRI and developed an accurate clinical decision diagnosis support model. With a model identified for clinical diagnosis decision support: in addition to aiding the radiologist with image classification, several common questions a patient may ask of a clinician can also be answered. The first question: what is the accuracy of this test at predicting malignancy? This question can be answered from the specificity calculation and we can say that there is 100% accuracy the model, when used with the dataset I have acquired, will identify a malignant tumor and aid in the decision

to recommend biopsy. The second question: if the test predicts malignancy, what is the confidence that a patient is truly malignant and not benign? This question can be answered using the PPV value, it can be said there is a 79% chance you have a malignant tumor. Put another way, there is only a 21% chance that you have been misdiagnosed as malignant and will have to undergo an un-necessary biopsy.

Image Misclassification

Although the model chosen has perfect malignant patient diagnosis for this dataset, misclassifications of malignant images still exist, and thus the issue of image misclassification should be better understood. In particular, an understanding of why some images in a patient's image set are misclassified while others in the set are not, is useful to fully understanding the clinical diagnosis support model proposed. To further our understanding, the dataset, and how it was created should be analyzed.

In the MRI in-bore guided biopsy process, images are used to identify possibly cancerous areas, regions of interest (ROI) that are considered abnormal compared to the surrounding prostate tissue. Obviously, the size of the suspected abnormal area will determine the number of images with an ROI for that particular patient. In the dataset used in this study, the number of images for a given patient ranges between 1 and 8, indicating various sized suspected tumor areas exist in the patient cohort used to compile the dataset.

The number of images acquired for a patient will determine the possible factors underlying patient's image misclassification. First, let us consider a patient with only one acquired image: it is possible that the biopsy area providing the ground truth label for this patient is not fully aligned with the image. This misalignment may cause the image to only capture a small area of the malignant tissue, with the rest of the image being benign. An example of a misaligned image is shown in Figure 24, where the majority of the ROI area is benign, shown in dark blue and light blue, or slightly malignant (Gleason score 6) shown in green.

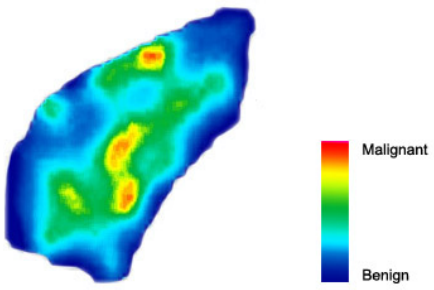

Figure 24. Heatmap of a misaligned ROI image.

The extracted features for such an image will skew a malignant labeled image towards those of a true benign labeled image. This skew, essentially a benign image labeled as malignant, can be considered noise in the dataset and will possibly confuse the classifier.

Tumor Heterogeneity

For malignant patients with multiple images, tumor heterogeneity can be a cause of mislabeled images in the patient's image set. Although the patient's biopsy shows a ground truth of malignant, and images in the patient dataset contain malignant features, because of the potential heterogeneity of the tumor, some images may contain most, or all, benign features. It is these benign images in the malignant patients dataset that will be mislabeled, and like the case of a single misaligned image, can be considered noise in the dataset.

Tumor heterogeneity, also referred to as genetic heterogeneity[19,57,58,59], describes both a spatial and temporal variation of differing cell populations in a tumor. Spatial heterogeneity of a tumor is caused by a varied stratification of cell types in a tumor [60,61], where as temporal heterogeneity, shows a variation of cell types in the tumor over time. In the context of this study, we are interested is the spatial heterogeneity of a tumor as it relates to image features for a given tumor ROI. The heterogeneous nature of cancer tumors has been observed in many cancer types including prostate cancer[62,63]. Intra-tumor heterogeneity can be the cause of inaccurate biopsy results when image guidance is not used. In the example in Figure 25, six horizontal MRI slices (shown by the black lines) have been used to identify an ROI in the prostate gland in which to perform a biopsy.

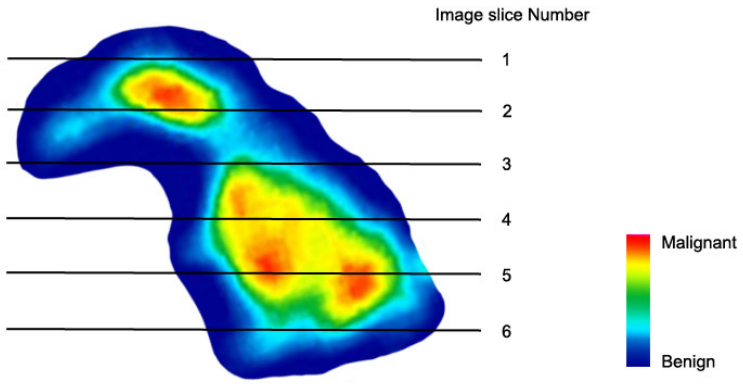

Figure 25. Prostate gland ROI identified by six MR image slices.

As shown in the color stratification in Figure 24, the ROI contains both benign tissue and various severities of malignant tissue. Assuming the Gleason score biopsy protocol of a primary grade pattern (Yellow), and a secondary grade pattern (Red) yields a score of 7 (3 + 4), the patient would be diagnosed as malignant. In the dataset used for this study, the six image slices shown in Figure 25 would all be labeled as malignant. Labeling all six images as malignant in this manner will most likely mislabel image slice 1, 3, and 6 as they contain little or no malignant tissue. An example of the tissue variation in horizontal image slices that may be present in a heterogeneous tumor such as that shown in Figure 25, and also possibly present in the dataset used in this study, is shown in Figure 26.

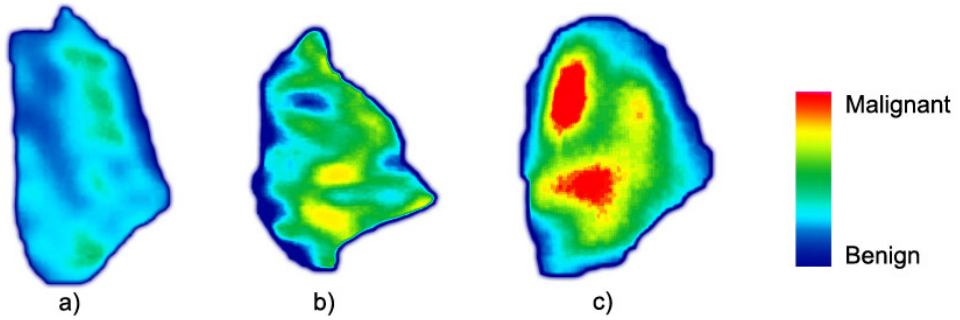

Figure 26. Image examples of malignant and benign tissue stratification in a heterogeneous tumor.

The image in Figure 26c shows a large amount of malignant tissue and extracted features would well represent the patterns seen in malignant tissue. On the other extreme, the image in Figure 26a shows little to no malignant tissue and the extracted features would represent those in a benign image. If these images are all from a patient diagnosed as malignant, the image in Figure 26a, and possibly Figure 26b would be mislabeled and add noise to the dataset. Added noise in the dataset causes weak predictor variables with low correlation to the class labels thus confusing the classifier during training and reducing classification accuracy.

The example in Figure 25 shows that image slices at either end of the ROI contain little to no malignant tissue suggesting that mislabeled images in image stack of a malignant patient may possibly be related to their position in the stack. An analysis of the misclassified images for all models used in this study shows no clear correlation to the spatial position in the image stack. In addition, analyzing misclassified malignant and benign images from the results of model testing shows no correlation to ROI size or the zone of the prostate gland where the image was acquired.

Although intra-tumor spatial heterogeneity can possibly cause mislabeling of images in malignant patients, it has been demonstrated in this study that the criteria of classifying a patient as malignant (if one or more images is classified as malignant the patient is classified as malignant), will overcome mislabeled images when diagnosing malignant patients. Of course, this criterion only applies if the dataset for a given patient contains multiple images. In some cases of smaller tumor size, acquiring multiple images may not be possible, so misclassification due to misalignment of image and tumor cannot be corrected for. A larger concern is the noise added to the dataset by mislabeled images causing classifier confusion and reduced classification accuracy. This classifier confusion may affect the classification of both benign and malignant images alike. Using the criteria set here for classifying a malignant patient, misclassification of malignant images can, to a degree, be corrected for, however, a single misclassified image from a benign patient will misclassify that patient as malignant reducing the benign classification accuracy and leading to over-biopsy.

Limitations

Several limitations have been identified for this study. As has been previously discussed, possible intra-tumor heterogeneity causing patient images to be mislabeled is one such limitation. A second limitation is the size and imbalance between benign and malignant image samples of the dataset being used to train the classification models. Small sized datasets in machine learning can cause overfitting of the models, specifically the autoencoder used for image de-noising and dimensionality reduction in this study. A small dataset also increases the negative effects of noise and outliers that may be present in the class labels and in the predictor variables. An imbalance in a dataset, such as the one used in this study, is expected since the radiologist who has determined which patients show likely malignant regions that should be biopsied strives for the best accuracy and least over-biopsy possible. It is interesting to note that if the radiologist were 100% accurate in predicting suspected malignant and benign regions of interest, we would have no benign images in the dataset to analyze. The result of an imbalanced dataset containing lower numbers of benign data samples verses malignant samples will cause the learner to be biased to the malignant class and will tend to misclassify benign labeled images. Although a k-fold cross-validation method was introduced to balance the dataset classes by reusing the benign images, this method is not as optimal as having a larger set of unique benign images. A method not used here to balance a dataset is to duplicate the smaller class data with added noise with a variance predicted from the variance of the features within the class. It may be of interest to test this method in the context of this study to determine if any classification accuracy would be achieved.

Lastly, a third limitation is the use of 2D images verses 3D images. Several studies[64,65,66] comparing the use of 2D and 3D MRI in cancer image analysis show that 3D images yield better classification accuracy over their 2D counterparts for many of the same feature extraction and classification methods used in this study, specifically LBP, which is used in the top performing model identified here. 3D image features are extracted in 3-dimensional voxels, which yield more information about the tissue being analyzed than can be obtained with traditional 2-dimensional feature extraction for a variety of cancer types, including prostate[67]. The 3D volumetric texture analysis method can overcome the issues associated with tumor heterogeneity, seen in the 2D image analysis methods used in this study, since more accurate features can be extracted over a volume in a 3D space of the tumor area.

Chapter 5: Conclusions

I have established an algorithm using extracted features and classification to enable computer image analysis of benign and malignant prostate cancer patients using MRI. The effectiveness of this algorithm achieves the overall goal of this project. From the results generated in establishing the computer image analysis algorithm, I have demonstrated that multiparametric MRI has better classification accuracy than using T2W or ADCmap images alone. While evaluating the individual models I recognized that an ensemble of classifiers would be beneficial, and to the best of my knowledge, introduced the ensemble of classifiers to mp-MRI prostate cancer computer image analysis. The ensemble of classifiers model was demonstrated to have a better classification accuracy than any of the individual models analyzed.

For this particular small sized and unbalanced dataset, I have shown the optimum patient level clinical diagnosis decision support model to be the mRMR feature selection method with the SVM classifier using a polynomial kernel. The optimum feature extraction method used with the optimum model is shown to be LBP with a bin size of 1. In identifying a machine learning method, which reduces the patient level malignant misdiagnosis to zero, I have achieved the specific aim of this project.

References

1. NIH: National Cancer Institute Cancer, Stat Facts: Prostate Cancer [Internet]. Available from: https://seer.cancer.gov/statfacts/html/prost.html
2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin. 2018; 68(1):7-30. doi: 10.3322/caac.21442
3. Banerjee, I, Hahn, L, Sonn G, Fan R, Rubin DL. Computerized multiparametric MR image analysis for prostate cancer aggressiveness-assessment [database on the Internet]. 2016 [cited 2018 May]. Available from: https://arxiv.org/abs/1612.00408
4. Liu L, Tian Z, Zhang Z, Fei B. Computer-aided detection of prostate cancer with MRI: technology and applications. Acad radiol. 2016;23(8):1024-1046. doi:10.1016/j.acra.2016.03.010
5. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017 Feb;542:115–118. doi:10.1038/nature21056
6. Stanford News: Stanford algorithm can diagnose pneumonia better than radiologists [Internet]. 2017 Nov 15. Available from: https://news.stanford.edu/2017/11/15/algorithm-outperforms-radiologists-diagnosing-pneumonia/
7. Ghai S, Haider MA. Multiparametric-MRI in diagnosis of prostate cancer. Indian J Urol. 2015;31(3):194-201. doi:10.4103/0970-1591.159606

8. Chung AG, Shafiee MJ, Kumar D, Khalvati F, Haider MA, Wong A. Discovery radiomics for multi-Parametric MRI prostate cancer detection. [database on the Internet]. 2015 [cited 2018 May]. Available from: https://arxiv.org/abs/1509.00111
9. Sarkar S, Das S. A Review of Imaging Methods for Prostate Cancer Detection. Biomed Eng Comput Biol. 2016;7(Suppl 1):1-15. doi:10.4137/BECB.S34255
10. Wang S, Burtt KE, Turkbey B, Choyke PM, Summers RM. Computer aided-diagnosis of prostate cancer on multiparametric MRI: A technical review of current research. Biomed Res Int. 2014;2014:789561. doi: 10.1155/2014/789561
11. American cancer society [Internet]. 2017 March 8, [cited 2017 May]. Available from: https://www.cancer.org/treatment/understanding-your-diagnosis/tests/understanding-your-pathology-report/prostate-pathology/prostate-cancer-pathology.html
12. Chen N, Zhou Q. The evolving Gleason grading system. Chin J Cancer Res. 2016;28(1):58-64. doi: 10.3978/j.issn.1000-9604.2016.02.04
13. Fusco R, Sansone M, Granata V, Setola SV, Petrillo A. A systematic review on multiparametric MR imaging in prostate cancer detection. Infect Agent Cancer. 2017;12:57. doi:10.1186/s13027-017-0168-z
14. Peng Y, Jiang Y, Yang C, Brown JB, Antic T, Sethi I, Schmidt-Tannawald C, et al. Quantitative analysis of multiparametric prostate MR images: differentiation between prostate cancer and normal tissue and correlation with gleason score: a computer-aided diagnosis development study. Radiology. 2013 Jun;267(3):787-96. doi: 10.1148/radiol.13121454. Epub 2013 Feb 7
15. Lemaître G, Martí, R, Freixenet J, Vilanova JC, Walker P, Meriaudeau F. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: a review. Comput Biol Med. 2015 May;60:8-31. doi: 10.1016/j.compbiomed.2015.02.009. Epub 2015 Feb 20.
16. Chan HP, Sahiner B, Helvie MA, Petrick N, Roubidoux MA, Wilson TE, Adler DD, Paramagul C, et al. Improvement of radiologists' characterization of mammographic masses by using computer aided diagnosis: an ROC study. Radiology. 1999;212(3):817-827
17. Dean JC, Ilvento CC. Improved cancer detection using computer-aided detection with diagnostic and screening mammography: prospective study of 104 cancers, AJR Am J Roentgenol. 2006;187(1):20-28
18. Hambrock T, Vos PC, Hulsbergen-van de Kaa CA, Barentsz JO, Huisman HJ. Prostate cancer: computer-aided diagnosis with multiparametric 3-T MR imaging-effect on observer performance. Radiology. 2013;266(2):521-530
19. Bai H, Lee AM, Yang L, Zhang P, Davatzikos C, Maris JM et al. Imaging genomics in cancer research: limitations and promises. Br J Radiol. 2016;89(1060):20151030. doi: 10.1259/bjr.20151030
20. UCSF Department of Radiology & Biomedical Imaging. Multi-Parametric Prostate Cancer Staging Exam [Internet]. Copyright 2018 The Regents of the University of California. Available at: https://radiology.ucsf.edu/patient-care/services/prostate-exam-study#accordion-dynamic-contrast-enhanced-dce-mri

21. Woodfield CA, Tung GA, Grand DJ, Pezzullo JA, Machan JT, Renzulli JF II. Diffusion-weighted MRI of peripheral zone prostate cancer: comparison of tumor apparent diffusion coefficient with gleason score and percentage of tumor on core biopsy. Am J Roentgenol. 2010 Apr;194(4):W316-22. doi: 10.2214/AJR.09.2651
22. Sankineni S, Osman M, Choyke PL. Functional MRI in Prostate Cancer Detection. BioMed Res Int. 2014;2014:590638. doi:10.1155/2014/590638
23. Liu S, Zheng H. Feng Y, Li W. Prostate cancer diagnosis using deep learning with 3D multiparametric MRI [database on the Internet]. 2017. aiXrv.org [cited 2018 June]. Available from: https://arxiv.org/abs/1703.04078
24. Wang X, Yang W, Weinreb J, et al. Searching for prostate cancer by fully automated magnetic resonance imaging classification: deep learning versus non-deep learning. Sci Rep. 2017;7:15415. doi:10.1038/s41598-017-15720-y
25. Yang X, Liu C, Wang Z, Yang J, Min HL, Wang L, Cheng KT. Co-trained convolutional neural networks for automated detection of prostate cancer in multi-parametric MRI. Med Image Anal. 2017 Dec;42:212-227. doi: 10.1016/j.media.2017.08.006. Epub 2017 Aug 24
26. Khalvati F, Wong A, Haider MA. Automated prostate cancer detection via comprehensive multi-parametric magnetic resonance imaging texture feature models. BMC Med Imaging. 2015 Aug;15:27. doi:10.1186/s12880-015-0069-9
27. Shah V, Turkbey B, Mani H, et al. Decision support system for localizing prostate cancer based on multiparametric magnetic resonance imaging. Med Phys. 2012 Jul;39(7):4093-103. doi: 10.1118/1.4722753.
28. Niaf E, Rouvière O, Mège-Lechevallier F, Bratan F, Lartizien C. Computer-aided diagnosis of prostate cancer in the peripheral zone using multiparametric MRI. Phys Med Biol. 2012 Jun 21;57(12):3833-51. doi: 10.1088/0031-9155/57/12/3833. Epub 2012 May 29
29. Morgan VA, S. Kyriazi S, Ashley SE, deSouza NM. Evaluation of the potential of diffusion-weighted imaging in prostate cancer detection, Acta Radiologica. 2009;48(6):695-703. doi: 10.1080/02841850701349257.
30. Maurer MH, Heverhagen JT. Diffusion weighted imaging of the prostate—principles, application, and advances. Transl Androl Urol. 2017;6(3):490-498. doi:10.21037/tau.2017.05.06
31. Ginsburg SB, Rusu M, Kurhanewicz J, Madabhushi A. Computer extracted texture features on T2w MRI to predict biochemical recurrence following radiation therapy for prostate cancer. Proc. of SPIE Vol. 9035, 903509 · © 2014 SPIE · CCC code: 1605-7422/14/$18. doi: 10.1117/12.2043937
32. Gudbjartsson H, Patz S. The rician distribution of noisy MRI data. Magn Reson Med. 1995 December;34(6): 910–914.
33. Nowak RD. Wavelet-based rician noise removal for magnetic resonance imaging. IEEE Trans Image Process. 1999;8(10):1408-19. doi: 10.1109/83.791966.

34. Kumar N, Nachamai M. Removal and filtering used in medical images. Orient J Comp Sci and Technol. 2017 Mar;10(1). doi: http://dx.doi.org/10.13005/ojcst/10.01.14.
35. Gonzalez RC, Woods RE. Digital Image Processing, 4th Addition, New York :Pearson; 2018:995
36. Im DJ, Ahn S, Memisevic R, Bengio Y. De-noising criterion for variational auto-encoding framework. Proc Conf AAAI Artif Intell [Internet]. 2017. Available from: http://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/download/14213/14374
37. Park E. Manifold learning with variational auto-encoder for medical image analysis. Semantic Scholar [Internet]. 2015. Available from: https://www.semanticscholar.org/paper/Manifold-Learning-with-Variational-Auto-encoder-for-Park/b7bbcd02597d797d359dc78b16fd154659bc1f6b
38. Jordan J. Variational autoencoders [Internet]. Available from: https://www.jeremyjordan.me/variational-autoencoders/
39. Gondara L. Medical image de-noising using convolutional de-noising autoencoders. 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, 2016, pp. 241-246. doi: 10.1109/ICDMW.2016.0041
40. Thibault G, Tudorica A, Afzal A, et al. DCE-MRI texture features for early prediction of breast cancer therapy response. Tomography. 2017;3(1):23-32. doi:10.18383/j.tom.2016.00241
41. Thibault G, Fertil B, Navarro C, Pereira S, Lévy N, Sequeira J, Mari J. Texture indexes and gray level size zone matrix application to cell nuclei classification. 10th International Conference on Pattern Recognition and Information Processing. 2009
42. Zwanenburg A, Leger S, Valli`eres M, L¨ock S. Image biomarker standardisation initiative [database on the Internet]. aiXrv.org 2016. Available from: https://arxiv.org/abs/1612.07003
43. Wibmer A, Hricak H, Gondo T, et al. Haralick Texture analysis of prostate MRI: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different gleason scores. Eur Radiol. 2015 Oct;25(10):2840-2850. doi:10.1007/s00330-015-3701-8
44. Haralick RM. Textural features for image classification. IEEE Trans Sys Man Cybern. 1973 Nov;SMC-3(6):610-621
45. Shakoor, MH. Lung nodule detection based on noise robust local binary pattern. Int J Sci Eng Res. 2014 May;5(5). ISSN 2229-5518
46. Ojala T, Pietikäinen M, Mäenpää T. Gray scale and rotation invariant texture classification with local binary patterns. In: Computer Vision - ECCV 2000. ECCV 2000. Lecture Notes in Computer Science, vol 1842. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45054-8_27
47. Tang X. Texture information in run-length matrices. IEEE Trans Image Process. 1998 Nov;7(11):1602-1609. doi: 10.1109/83.725367

48. Annavarapu JK. Statistical feature selection for image texture analysis. Int J Res Eng Technol. 2015 Aug;(2)5:546-550
49. Al-Kadi, O. A fractal dimension based optimal wavelet packet analysis technique for classification of meningioma brain tumors. Conf Proc IEEE Int Conf Image Process. 2009:4177-4180. doi: 10.1109/ICIP.2009.5414534
50. Lin HT. Texture classification using fractal-based features and support vector machines [Internet]. Available from: http://www.work.caltech.edu/~htlin/course/doc/FractalFinal.pdf
51. Sarkar N, Chaudhuri BB. An efficient differential box counting approach to compute fractal dimension of images. IEEE Trans Syst Man Cybern. 1994 Jan;24(1):115-120. doi:10.1109/21.259692
52. Farnia F, Kazerouni A, Babveyh A. Information-based feature selection [Internet]. Stanford University. Available from: http://cs229.stanford.edu/proj2014/Farzan%20Farnia,%20Abbas%20Kazerouni,%20Afshin%20Babveyh,%20Information%20based%20feature%20selection.pdf
53. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Statist Soc B. 2005:67(2):301-320. 1369-7412/05/67301
54. Hamiane M, Salman FS. MRI brain image analysis and classification for computer-assisted diagnosis. Int J Econ and Manage Sys. 2017;2:229-236.
55. Bahadure NB, Ray AK, Thethi AP. Image analysis for MRI based brain tumor detection and feature extraction using biologically inspired BWT and SVM. Int J Biomed Imaging. 2017;2017:1-12. doi:10.1155/2017/9749108
56. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. New York: Springer; 2009:423
57. Zhou M, Leung A, Echegaray S, Gentles A, Shrager JB, Jensen KC, et. al. Non-small cell lung cancer radiogenomics map identifies relationships between molecular and imaging phenotypes with prognostic implications. Radiology. 2017;286(1):307-315. doi: 10.1016/j.lungcan.2017.10.015 PMCID: PMC5749594
58. Sala E, Mema E, Himoto Y, Veeraraghavan H, Brenton JD, Snyder A, Weigelt B, et al. Unraveling tumor heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging. Clin Radiol. 2017;72(1):3–10. doi:10.1016/j.crad.2016.09.013
59. Davnall F, Yip CS, Ljungqvist G, Selmi M, Ng F, Sanghera B, et al. Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice. Insights Imaging. 2012 Dec;3(6):573-89. doi: 10.1007/s13244-012-0196-6. Epub 2012 Oct 24.
60. Samuel N, Hudson TJ. Translating genomics to the clinic: implications of cancer heterogeneity. Clin Chem. 2013; 59(1):127-37. doi: 10.1373/clinchem.2012.
61. O'Connor JP, Rose CJ, Waterton JC, Carano RA, Parker GJ, Jackson A. Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. Clin Cancer Res. 2015 Jan 15;21(2):249-57. doi: 10.1158/1078-0432.CCR-14-0990. Epub 2014 Nov 24.

62. Stoyanova R, Takhar M, Tschudi Y, Ford JC, Solórzano G, Erho N, et al. Prostate cancer radiomics and the promise of radiogenomics. Transl Cancer Res. 2016;5(4):432-447. PMCID: PMC5703221
63. Tosoian JJ, Antonarakis ES. Molecular heterogeneity of localized prostate cancer: more different than alike. Transl Cancer Res. 2017;6(1):47-50.
64. Ortiz-Ramon R, Larroza A, Arana E, Moratal D. A radiomics evaluation of 2D and 3D MRI texture features to classify brain metastases from lung cancer and melanoma. Conf Proc IEEE Eng Med Biol Soc. 2017 Jul;493-496. doi: 10.1109/EMBC.2017.8036869.
65. Chen W, Giger ML, Li H, Bick U, Newstead GM. Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images. Magn Reson Med. 2007 Sep;58(3):562-71. PMID:17763361. doi: 10.1002/mrm.21347.
66. Arai K, Herdiyeni Y, Okumura H. Comparison of 2D and 3D local binary pattern in lung cancer diagnosis. Int J Adv Comp Sci Appl. 2012;3(4):89-95. doi: 10.14569/IJACSA.2012.030416
67. Depeursinge A, Foncubierta-Rodriguez A, Van De Ville D, Müller H. Three-dimensional solid texture analysis in biomedical imaging: review and opportunities. Med Image Anal. 2014 Jan;18(1):176-96. doi: 10.1016/j.media.2013.10.005. Epub 2013 Oct 22.